Deciphering the mtDNA record of prehistoric population movements in Oceania

A thesis submitted in partial fulfilment of the
requirements for the degree of
Doctor of Philosophy in Zoology

by
Melanie J. Pierson

University of Canterbury
2007

ABSTRACT

This thesis uses mitochondrial DNA (mtDNA) phylogenies to explore patterns of past human mobility in Oceania. To extend the current knowledge of mtDNA variation in Oceania, 20 entire mt genomes were sequenced and analysed in a data set of more than 144 sequences from Australia, Oceania, Island Southeast Asia and Taiwan. The MinMax Squeeze method enabled this large data set to be analysed with an optimality criterion (Pierson *et al.* 2006). The analysis revealed two major groups of haplogroups in Oceania, distinguished by the relationships to others outside of the region: an 'ancient' set of types whose phylogenies and distributions suggest they are descended from the Pleistocene-era settlers of Near Oceania, and a second 'young' group whose presence in Oceanic populations may reflect more recent movements into Near Oceania. The detailed phylogenies of these haplogroups presented here will aid in future investigations of human mtDNA in Oceania, allowing samples to be screened by defining mutations to target haplogroups of interest.

A large data set of global entire human mt DNA sequences was assembled from public data bases and tested for evidence of selection and recombination. These tests, and phylogenetic analyses of random subsets of the data set, found high levels of homoplasy in the sequences. Homoplasy in the control region of the mtDNA molecule was examined in particular, resulting in a relative scale of mutability at each position of the ~1kb sequence. Subsequent phylogenetic tests of weighting schemes derived from this analysis for the control hypervariable region I (HVR-I) did not show demonstrable improvements over the unweighted examples, but did highlight instances in which the HVR-I sequence failed to predict the more robust trees generated by the coding region.

Finally, the HVR-I and diagnostic SNPs were sequenced in a set of 46 Polynesian samples from Auckland, and this data was analysed within a large set of HVR-I sequences (>4000) from Oceanic, Asian and the American populations available from public data bases. These analyses were informed by the whole mtDNA phylogenies generated earlier in the project, and add population level data to the emerging picture of prehistoric female mobility gained from entire mtDNA analyses.

Contents

Abstract iii

List of Tables x

List of Figures xi

Acknowledgements xii

Chapter 1 Introduction 1

1.1 The Pacific and its peoples 1

1.2 Pacific prehistory: evidence from archaeology and historical linguistics 3

1.3 Genetic markers and inferences for prehistory 6

Commensal organisms 7

Y chromosome analyses 8

Mitochondrial DNA analyses 9

1.4 Thesis outline 10

Chapter 2 Analysis of Oceanic data set 13

2.1 Collection of new sequences 13

Sample sources 13

Amplification, sequencing and assembly of mt genomes 14

2.2 Data set details and phylogenetic methods 16

Data set details 16

Maximum parsimony, the MinMax Squeeze and consensus networks 16

2.4 Results 18

Oceania-127 18

Oceanic-133 19

2.5 Discussion 22

Geographic distribution of haplogroups 22

Monophyly of previously described M haplogroups 23

The MMS analysis 23

Chapter 3 mtDNA Haplogroups in Oceania 27

3.1 Methods 27

Phylogenetic analysis 27

Molecular dating 29
3.2 Results and discussion 29
M/Q and M/M29 haplogroups 30
N/R/P, analysed with N/R/R21 33
M27 and M28 haplogroups 36
N/R/B4a haplogroup 38
M/M7bc, analysed with M22 41
N/R/B5a haplogroup 44
3.3 Issues in dating human mtDNA phylogenies 45

Chapter 4 Variation in human mtDNA 51
4.1 Introduction 51
Structure and function of mtDNA 51
Transcription 54
Replication and repair 55
mtDNA inheritance 55
mtDNA recombination 57
Hypervariable sites and selection 58
4.2 Assembly of the whole mtDNA data sets and haplogroup assignment 60
4.3 Variation in the global data set 64
Average pairwise distance between haplogroups. 64
Variability by mtDNA region. 65
4.4 Tests of selection. 68
4.5 Testing for recombination. 70
4.6 Discussion. 73

Chapter 5 Phylogenetic analysis of homoplasy in the global data set 79
5.1 Methods used in 75-taxa analysis 79
Construction of the random data sets 79
PAUP* and MMS analysis 80
5.2 Results of the 75-taxa analysis 81
Homoplasy in the coding region 81
Mutation hotspots and conserved areas in the control region 85
5.3 Methods used for the coding vs. HVR-I analysis 89

Preparation of data sets 89
PAUP* parsimony analysis 90
5.4 Results of the coding vs. HVR-I analysis 91
5.5 Discussion 94

Chapter 6 Collection and analysis of new Polynesian HVR-I samples in an Oceanic context 97
6.1 Collection of HVR-I and coding SNP data from 46 Polynesian samples 97
Sample collection 97
mtDNA sequencing 97
Results and discussion 98
6.2 Compilation of HVR-I data sets from Oceania, Asia and the Americas 100
6.3 The distribution of HVR-I haplotypes in Oceania 102
6.4 Phylogenies from HVR-I sequences: N/R/P1, M/Q1 and N/R/B4a 110
N/R/P1 110
M/Q 112
N/R/B4a 112
6.3 Discussion 115

Chapter 7 Conclusions and Future Directions 117

References 121

Appendices 137
A. Pierson et al (2006) 137
B. Annotated reference sequence 147
C. Methodology details 157
C2.1 Molecular methodology details 158
C5.1 PAUP* commands 75-taxon analysis 161
C5.2 MMS parameters 75-taxon analysis 162
C5.3 C++ code for random selection of 15 taxa from globalhapscoded.nex 163
C5.4 PAUP* commands for coding vs. HVR-I analysis 164
D. Supplementary Figures 165
D3.1 N/W consensus network and labelled phylogeny 166
D3.2 N/S consensus network and labelled phylogeny 167

D3.3 N/R/B4b consensus network and labelled phylogeny 168
D3.4 L1c consensus network and labelled phylogeny 169
D3.5.East Asian skeleton phylogeny macrohaplogroup N (Kong et al 2006) 170
D3.6 East Asian skeleton phylogeny macrohaplogroup M (Kong et al 2006) 171
D4.1 Haplotype tree 172
E. Supplementary Tables 185
E1.1 Oceanic population size estimates 1930-2003 186
E1.2 Y chromosome review table 187
E1.3 Oceanic mtDNA studies review table 190
E2.1 Polymorphism lists for mt genome sequences from this study 193
E5.1 Parsimony scores for characters from 75-taxon analysis 200
E6.1 Haplotype details HVR-I nt16065-nt16373 data set (Oceanic samples) 209
E6.2 Haplotype details HVR-I nt16189-nt16373 data set 215
F. Digital appendices ..... CD-ROM
F2.1 Oceanic-133.nex
F3.1 Haplogroup nexus files folder
F3.1.1_QM29.nex
F3.1.2_P_R21_R12.nex
F3.1.3_M27_M28.nex
F3.1.4_B4a.nex
F3.1.5_M7bc_nex
F3.1.6_W.nex
F3.1.7_S.nex
F3.1.8_B4b.nex
F3.1.9_L1c.nex
F3.1.10_B5a.nex
F3.2 Folder of haplogroup networks and labelled tree figures
F3.2.1_QM29.pdf
F3.2.2_P_R21_R12.pdf
F3.2.3_M27_M28.pdf
F3.2.4_B4a.pdf
F3.2.5_M7bc_pdf
F3.2.6_W.pdf
F3.2.7_S.pdf
F3.2.8_B4b.pdf
F3.2.9_L1c.pdf
F3.2.10_B5a.nex
F4.1 Global dataset folder
F4.1.1_globalmtDNA.nex

F4.1.2_globalmtDNAhaps.nex

F4.1.3_globalmtDNAcomplete.nex

F4.1.4_globalmtDNAhapscomplete.nex

F4.1.5_globalmtDNAhaps1.meg

F4.1.6_globalmtDNAhaps2.meg

F4.1.7_globalmtdataset.xls

F5.1_globalhapcoded.nex

F6.1 HVR-I analyses folder

F6.1.1_HVRI16065.arp.txt

F6.1.2_HVRI16189.arp.txt

F6.1.3_HVRIanalyses.xls

List of Tables

Table 1.1 Distribution of major Y-chromosome haplogroups in Oceania and surrounding regions..... 9
Table 2.1 New sample details 14
Table 2.2 Oceanic data set sequences 17
Table 3.1 Haplogroup data set details 30
Table 3.2 TMRCA estimates 46
Table 3.3 Ratios of rho for vertices with more than 10 descendants 47
Table 3.4 Coding-region substitution rates from Atkinson (2006) 48
Table 4.1 Global data set details 61
Table 4.2 Data set file details 62
Table 4.3 Global data set haplogroup details 63
Table 4.4 Variability by mtDNA region 66
Table 4.5 Variability in tRNA genes 67
Table 4.6 Tests of selection: Tajima's *D* and McDonald-Kreitman results 69
Table 4.7 Phi test for recombination 71
Table 5.1 'Hotspots' in the coding region 82
Table 5.2 Haplogroups used in coding region vs. HVR-I phylogeny comparison 89
Table 5.3 Coding region vs. HVR-I phylogeny comparison: haplogroup results 93
Table 5.4 Homoplastic bases in Random_30 data set and 75-taxa steps 98
Table 6.1 Auckland sample set details and results 101
Table 6.2 HVR-I nt16065-16373 data set accession details 103
Table 6.3 HVR-I nt16065-16373 data set geographic details 104
Table 6.4 HVR-I nt16189-16370 data set geographic details 105
Table 6.5 HVR-I nt16065-16373 data set diversity summary results 105
Table 6.6 HVR-I haplotypes nt16189-16370 data set diversity summary results 107
Table 6.7 HVR-I nt16065-16373 data set haplotypes found in five or more regions 108
Table 6.8 HVR-I haplogroup defining polymorphisms (nt16189-nt16370 data set) 108

List of Figures

Figure 1.1 Oceania general reference map 2
Figure 1.2 Lapita potsherds from Nukuleka, Tonga 4
Figure 2.1 Sample locations 13
Figure 2.2 Contig example PAI9 (DQ372869) 15
Figure 2.3 Oceanic-127 consensus networks 20
Figure 2.4 Oceanic-133 consensus networks 21
Figure 3.1 Q haplogroup with M29 consensus tree and branch-labelled phylogeny 31
Figure 3.2 P haplogroup with R21 and R12 consensus network and branch-labelled phylogeny 34
Figure 3.3 M27 and M28 haplogroup minimal tree and labelled phylogeny 37
Figure 3.4 B4a consensus network and branch-labelled phylogeny 40
Figure 3.5 M/M7bc with M22 consensus network and labelled phylogeny 42
Figure 3.6 B5a haplogroup minimal tree and labelled phylogeny 44
Figure 4.1 Mitochondrial energy production and mtDNA 52
Figure 4.2 mtDNA transcription and replication features 53
Figure 4.3 Intra-individual mtDNA recombination 58
Figure 4.4 Absolute pairwise distances (coding region) between groups in the global dataset 65
Figure 4.5 Nucleotide and amino acid variability chart 66
Figure 4.6 Average Ka/(Ks+constant) values 70
Figure 4.7 Consensus network of most parsimonious trees, Kraytsberg et al (2004) recombinants 72
Figure 4.8 Consensus network of most parsimonious trees, Random_30 dataset entire mtDNA 74
Figure 5.1 Distribution of homoplasious bases in protein-coding genes 83
Figure 5.2 Chart of control region 'hotspots' 86
Figure 5.3 Example of parsimony haplogroup scoring for coding region vs HVR-I comparison 90
Figure 5.4 Haplogroup steps required for trees coding vs HVR-I comparisons 92
Figure 6.1 Distribution of HVR-I haplotypes (nt16189-nt16370) in Oceania 109
Figure 6.2 N/R/P1 HVR-I haplotypes in Oceania 111
Figure 6.3 M/Q1 HVR-I haplotypes in Oceania 113
Figure 6.4 N/R/B4a HVR-I haplotypes in Oceania 114

ACKNOWLEDGMENTS

Firstly I would like to thank Neil Gemmell for much appreciated advice and guidance throughout the course of this project. Many thanks are also due to David Penny, Mike Steel, Barbara Holland, Rosa Martinez for their assistance and support. John Clegg, Wulf Schiefenhovel, Matthew Hurles and Brad Fris provided samples for this project, and the project was supported by a grant from the Marsden Fund, and a scholarship from the Allan Wilson Centre for Molecular Ecology and Evolution. Many people provided excellent technical assistance, and I would like to thank Trish McLenachan, Katherine McBride, Jan McKenzie, and Gavin Robinson in particular for their help, Iris Vargas-Jentzsch and David Bryant for last-minute computing assistance, and Tamsin Braisher for proof-reading parts of this thesis.

All of the members of the University of Canterbury Molecular Ecology Laboratory and the AWC project three deserve grateful thanks for their support and friendship over the past four years. For many valuable discussions of Pacific prehistory and genetics, financial help and patience, I would like to thank Lisa Matisoo-Smith; and Andreas and Julie Matisoo, for providing me with an ideal writing environment for the final completion of this project.

Finally, I would like to thank my family, and Alastair, Alice, Andrew, Barbara, Brent, David, Hayley, Kath, Kim, Margee, Maxine and Rachel for their support, advice and great company over the past few years.

1. Introduction

This chapter introduces the context for the analyses undertaken over the course of this project which aimed to elucidate pathways of human settlement and post-settlement interactions in the Pacific using mtDNA as a marker. The current understanding of prehistory of the Pacific derived from the fields of archaeology and historical linguistics is briefly reviewed, along with the contribution genetic studies have made to the current knowledge, before the structure of this thesis is outlined.

1.1 The Pacific and its peoples

The Pacific Ocean covers more than a third of the earth's surface and is home to more than twenty thousand islands. Combined, these islands form approximately 1.3 million square kilometres of land, making up only ~0.7% of the total area of the Pacific (Terrell 1986). About 70% of the land mass is represented by New Guinea, and a further 20% by the islands of New Zealand. In the mid-eighteenth century, Charles de Brosse, a French historian and geographer named the entire Pacific Island region Polynesia (from Greek *poly* many and *nesos* island). In the early 1830s the French explorer Jules Sebastien César Dumont d'Urville subdivided this to three groups, introducing the terms Melanesia and Micronesia (*melas*, black; *micros*, small, Fischer 2002).

Melanesia includes the large island of New Guinea (politically Irian Jaya and Papua New Guinea) and the subequatorial islands of the southwest Pacific: the offshore island groups of New Guinea, the Bismarck Archipelago, the Solomon Islands, the Santa Cruz Islands, the Banks Islands, Vanuatu, New Caledonia and Fiji (Figure 1.1). The Micronesian islands to the north of Melanesia (Palau, the Marianas, the Federated States of Micronesia, the Marshalls, Nauru and Kiribati) are situated east of the Philippines and lie mostly north of the equator. Polynesia is often described as a triangle, with apices at Hawai'i in the North Pacific, New Zealand to the south, and Easter Island to the east.

While useful geographical boundaries, Polynesia, Melanesia and Micronesia do not necessarily describe homogenous population groupings, although in the past they have been used to define population 'types' (Bellwood 1978, Howells 1973). The people of Polynesia show homogeneity in physical appearance, languages and cultures but this is not the case for Micronesian and Melanesian peoples. Melanesia in particular has a remarkable diversity of human cultures and languages between and within islands and island groups (Green 1989).

All of the Polynesian languages are members of the Oceanic subgroup of the Austronesian family, and the languages

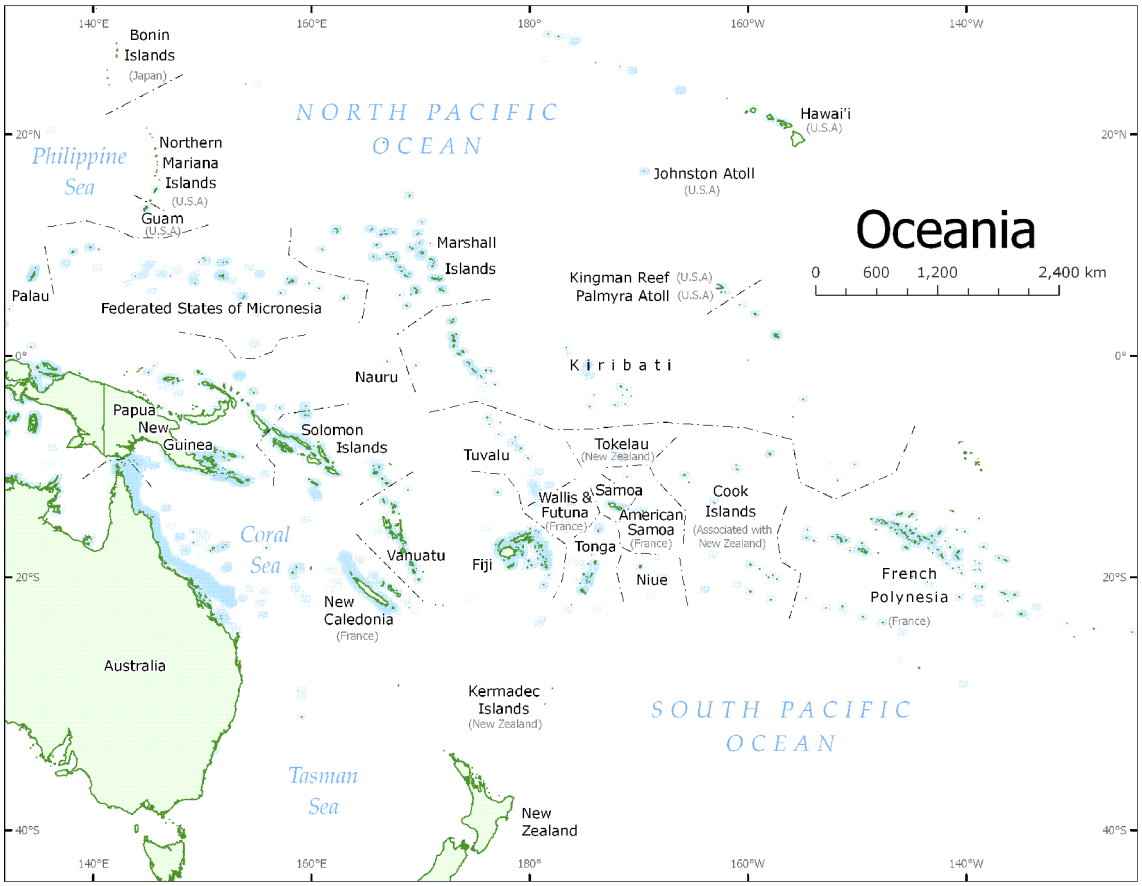


Figure 1.1 Oceania general reference map. From the East-West Center Spatial Information Services, Honolulu, Hawai'i http://www2.eastwestcenter.org/environment/spatial/ewc_sdi/

spoken in Micronesia are also Austronesian. This large family of languages is spoken throughout the Indo-Malaysian archipelago, Taiwan, the Philippines and Pacific and in pockets in mainland Southeast Asia. It was probably the most widespread language family in the world before AD1500, with speakers from Madagascar to Easter Island. However, in Melanesia, in addition to Austronesian languages, about 700 non-Austronesian languages are spoken. As many of these are from New Guinea they are often grouped together as 'Papuan', but this can be deceptive as they belong to at least 12 different language families (Kirch 2000).

The largest populations in Oceania are found in Papua New Guinea (~5.5 million) and New Zealand (~4 million). Demographic estimates for several countries and territories within the Pacific (Appendix E1.1) dating back to 1930 show marked increases in population size in most island groups over the past 70 years. The extent of immigration into the Pacific during the historic period differs between different islands and island groups. For example New Zealand has a large proportion of people of non-Maori descent (6 in 7

people or ~86%, Statistics New Zealand, http://www.stats.govt.nz/products-and-services/Articles/census-snpsht-maori-Apr02.htm; accessed 20/09/07). In the Solomon Islands 93% of the ~470,000 Solomon Islanders identify as Melanesian (2002 census), while in Fiji 54.3% of ~840,000 identify as Fijian (2004 census), and only 37% of Guam's population of ~170,000 identified as Chamorro in a 2000 census (Encyclopaedia Britannica® Online Academic Edition, © 2006 Encyclopaedia Britannica, Inc.).

There are several recorded historic instances of population decreases in the Pacific following European contact. For example in Hawai'i an estimated 30% of Native Hawaiian men left to work in North America between 1820 and 1840, with less than 15% returning (Cann and Lum 2004). The island of Rapa in the Austral Islands in Eastern Polynesia suffered an extreme population crash in the 19th century dropping to a low of 120 individuals within 50 years of contact, due to introduced disease and severe storms, from an estimated 2400 inhabitants when first encountered by Europeans (Martinson *et al.* 1993, Hurles *et al.* 2003).

A second descriptive division of the Pacific Islands distinguishes Near Oceania from Remote Oceania (Pawley and Green 1973). This grouping takes into account biogeographic factors and is of greater relevance to prehistoric research in the Pacific than the tripartite divisions of Melanesia, Polynesia and Micronesia. Near Oceania includes all the islands to the west of the Solomon Islands, which have relatively small inter-island distances. The islands of Near Oceania are intervisible and the plants and animals more diverse. Remote Oceania includes all of the islands of Polynesia and Micronesia, and those of eastern Melanesia that are separated from Near Oceanic islands by water gaps of more than 350 kilometres. Voyaging to and between the islands of Remote Oceania requires technically advanced long-distance sailing technologies, and adaptation to a more limited set of resources (Green 1991). The initial movements into Remote Oceania mark a key point in Oceanic prehistory.

1.2 Pacific prehistory: evidence from archaeology and historical linguistics

Historically much attention has focused on determining the origins of the Oceanic peoples. For example a century ago John Macmillan Brown argued for the initial settlement of the Pacific Islands by foot across land bridges, with some rafting across short sea-stretches, thousands of years ago by a palaeolithic, Caucasian people who were superseded within the last ten thousand years by another Caucasian race, who were seafarers. This was only a partial replacement; only men were thought capable of the journey as 'a few hundred miles of sea were sure to daunt primitive women from venturing her children and her household gods upon so dangerous an element; the thousands of miles between resting places in Polynesia made such ventures impossible for them' quoted in Howe (1999:318). The many different historical theories of origins,

and current academic and alternative models of Pacific settlement are reviewed and discussed in Howe (2003).

Archaeological studies indicate the first settlers of Near Oceania reached New Guinea during the Pleistocene by at least 40 000BP, and the Solomon Islands to the east by about 30 000BP (O'Connell and Allen 2004). Lower sea levels during the Pleistocene left a large landmass linking present-day Tasmania, the Australian mainland and New Guinea, known as Sahul. Early settlers had to make sea-crossings to reach Sahul from the west, where present-day Island Southeast Asia was part of a larger landmass. Technology for sea-crossings was also required to leave Sahul and reach the islands of New Britain and the Solomon Island chain (Kirch 2000).

In sharp contrast to the Pleistocene era settlement of Near Oceania the earliest evidence of settlement of the islands of Remote Oceania dates to approximately 3200BP (Kirch 2000). Within a period of only a few hundred years between about 3400BP and 3100BP sites appear in the archaeological record in New Caledonia, the Bismarcks, the Solomons, Vanuatu, Fiji, Samoa and Tonga with similar assemblages including distinctive pottery, along with evidence of a mixed horticultural and maritime subsistence. This archaeological horizon has become known as the 'Lapita Cultural Complex', after a site on the coast of New Caledonia excavated by Gifford and Shutler in the early 1950s (Kirch 2000). One of the most distinctive features of the Lapita Cultural Complex is the decorated pottery often recovered from sites (Figure 1.2).

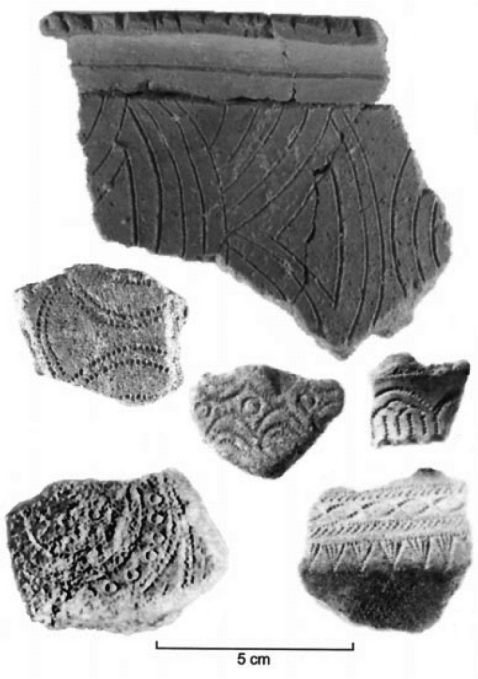


Figure 1.2 Lapita potsherds from Nukuleka, Tonga. From Burley and Dickinson 2001, PNAS 98:11830. Copyright 2001 National Academy of Sciences, U.S.A.

It is thought that a proto-Polynesian society developed after the Lapita settlement of western Polynesia, in Fiji, Samoa and Tonga, and subsequent Polynesian explorers journeyed from this 'homeland' out to the eastern Pacific (Kirch 2000). There is evidence the first inhabitants of western Polynesia maintained links with Lapita communities to the west; seen for example in ceramic features shared between early Tongan sites and sites in the Santa Cruz Islands (Burley and Dickinson 2001).

The Marquesas Islands were reached by approximately 300AD, and within about 1000 years of this, the many islands of the Polynesian 'triangle' had been settled (Kirch 2000). The earliest sites from the islands of Palau in the western part of Micronesia indicate earlier settlement; by about 3500BP (Liston 2005). Central and eastern Micronesia are thought to have been colonized later, between about 2000BP and 1500BP and evidence from historical linguistics and similarities of pottery form suggest that these later settlers were from a Lapita culture (Kirch 2000).

The distribution of Austronesian languages in Oceania, and the similarities and differences seen between them, play an important role in theories of prehistory of the Pacific. Robert Blust has compiled an extensive data set of the 1200 Austronesian languages which are subdivided to 10 main subgroups. Nine of these are situated in Taiwan, suggesting that the tenth group of Austronesian languages spread out into the Pacific region from Taiwan (Diamond 2000). Phylogenetic analysis of Austronesian language features has found strong support for the model of Austronesian expansion proposed by the traditional methods of historical linguistics (Gray and Jordan 2000), with a step-wise progression from Taiwan south and southeast into Oceania. The lack of cognates within the non-Austronesian languages is consistent with a greater time depth for separation than the Austronesian languages in the region. Dunn *et al.* (2005) have analysed non-Austronesian languages from Island Melanesia using structural features rather than cognates as characters and found the resulting patterns to be complex, but with large-scale genealogical clustering (Gray 2005).

Green (2003) has reviewed orthodox models for the development of the Lapita cultural complex in Near Oceania; summarizing the many theories put forward into four categories. The first set is named the 'Express Train to Polynesia' (ETP) group and Green states (2003:6): 'the metaphor of a train journey, with few or no stops of any duration along the way, conflicts with all the available evidence'. While formulated initially on the basis of archaeological and historical linguistic evidence by Peter Bellwood (1978) this model is no longer widely held by archaeologists yet is often presented as the orthodox model for testing in molecular studies (for example Capelli *et al.* 2001, Kayser *et al.* 2006). Green writes (2003:5) 'in continuing to test and support the ETP model, some molecular biologists are doing a great disservice to many. On a topic of mutual

concern to interdisciplinary researchers, they are not taking enough care in reading the work by colleagues in other fields'.

The second set of models is called the 'Bismarck Archipelago Indigenous Inhabitants' (BAII) and proposes the Lapita sites are representative of long-term continuity through time in the Bismarck Archipelago, and could have developed without incomers (Gosden 1992, White *et al.* 1988). The third set 'Slow Boat to the Bismarcks (SBB) emphasizes interactions along a 'voyaging corridor' from Indonesia through the Bismarcks, and onwards to the Solomon Islands. This sources the Lapita settlers, and in some formulations, the expansion of the Austronesian language family, in Island Southeast Asia (Terrell 2004, Oppenheimer 2004).

Green's fourth set of models, the 'Voyaging Corridor Triple I' (VC Triple I) extends earlier versions of the Intrusion, Integration and Innovation (Triple-I) model of Green (1991), to incorporate the voyaging corridor concept of the third model set. The Triple-I model describes the appearance of the Lapita Cultural Complex in Near Oceania as evidence of a migration of people, but also places emphasis on the integration of these new settlers with existing populations, and the *in situ* development of the full suite of Lapita characteristics within Near Oceania.

Kirch and Green (1987, 2001) advocate a phylogenetic approach to reconstructing the world of the ancestral Polynesians, drawing together evidence from biological anthropology, archaeology, historical linguistics and cultural ethnology. This use of the phylogenetic model in historical anthropology 'emphasizes historical sequences of cultural differentiation or divergence within related groups, regardless of the mechanism of transmission' (Kirch and Green 2001:13). An important aspect of this method is taking the evidence from the subdisciplines independently:

> 'historical linguistics, archaeology, comparative ethnology, and biological anthropology independently contribute their data and assessments to the common objective of historical reconstruction...the analytical power of the triangulation method and the robustness of the historical reconstructions derived from it only holds, however, if one treats each data source separately, respecting the relevant subdisciplinary methods, inferences, and conclusions as they are developed independently, based exclusively on the evidence from that field' (Kirch and Green 2001:42-43).

1.3 Genetic markers and inferences for prehistory

Studies of human genetic diversity in Oceania and neighbouring regions have tended to concentrate on the

uniparentally inherited mtDNA and Y-chromosome markers, although there have been studies of nuclear DNA (Hagelberg *et al.* 1999a, Lum *et al.* 1998, Martinson *et al.* 1993, Lie *et al.* 2007). Below the increasing contribution to our understanding of Pacific prehistory gained from genetic studies of commensal organisms is briefly discussed, and Y-chromosomal and mitochondrial DNA studies in this region reviewed.

Commensal organisms

Analyses of the patterns of genetic variation within commensal organisms that travelled with people as they moved around the Pacific are making a strong contribution to the understanding of Oceanic prehistory. Matisoo-Smith *et al.* (Matisoo-Smith *et al.*. 1998, Matisoo-Smith and Robins 2004) have demonstrated the value of this approach through analyses of mtDNA variation in Pacific rat (*Rattus exulans)* populations in the Pacific. Their results suggested multiple post-introduction contact events between east Polynesian island groups, even those as geographically distinct as Hawaii and New Zealand. mtDNA sequences from *R. exulans* bones from archaeological contexts in New Zealand also support repeated contacts between New Zealand, where the date of introduction of *R. exulans* remains controversial (Holdaway 1996, Wilmshurst and Higham 2004), and other parts of Oceania (Matisoo-Smith 2002).

A large study of *R. exulans* mtDNA samples from throughout Island Southeast Asia and Oceania found haplotypes grouped in three distinct clusters, with the sequences from Remote Oceania shared only by a small number from Halmahera, while those from Near Oceania were grouped with samples from Southeast Asia as well (Matisoo-Smith and Robins 2004). This suggests that the rats taken into Remote Oceania were also intrusive to Near Oceania.

The use of commensal animals as models of human pathways has been continued by Matisoo-Smith's group, with work on dog, chicken and pig variation (Matisoo-Smith 2002, Larson *et al.* 2007, Storey *et al.* 2007). Ancient DNA from Polynesian dog bones has been included in a larger study concentrating on the origins of the Australian dingo. The 19 mtDNA control region samples from the Cook Islands, New Zealand and Hawai'i belonged to two haplotypes, one of which is widespread amongst dogs from Asia and the Americas, but the other found only in two samples from Indonesia (Savolainen *et al.*. 2004).

Plants introduced to Remote Oceania by humans also show great potential to add a further dimension to prehistoric studies. Clarke *et al.* (2006) have examined chloroplast and nuclear markers in the Polynesian bottle gourd, concluding that there is evidence of shared ancestry with both the American subspecies, believed to have reached the Americas from Asia in the early Holocene (Erickson *et al.*. 2005), and the Asian subspecies.

Storey *et al.* (2007) have recovered ancient mtDNA from a chicken bone found in a pre-Columbian archaeological site in south central Chile. This was found to have an identical sequence to archaeological chicken remains from Samoa and Tonga, providing further evidence of voyaging between Polynesia and the Americas.

Y-chromosome analyses

Knowledge of variation in the male-specific (non-recombining) region of the Y chromosome (the MSY or NRY), comprising 95% of the chromosome's length, has greatly increased in recent years with the complete sequence for an individual completed in 2003 (Skaletsky *et al.*. 2003). Large-scale studies of Y-chromosome variation have demonstrated its utility in examining human population history (Jobling and Tyler-Smith 2003, Ke *et al.*. 2001, Underhill *et al.*. 2000).

The first papers on Y-chromosome diversity in Oceanic populations appeared in the 1990s (Hagelberg *et al.*. 1999b, Hurles *et al.*. 1998, Spurdle *et al.*. 1994), but as few markers were known at the time their power to distinguish haplotypes was not great. The increasing number of markers available over the past decade provides a much finer resolution to recent studies (for example Cox and Lahr 2006, Hurles *et al.*. 2005, Kayser *et al.* 2006). Several analyses have highlighted apparent differences between the patterns of Y-chromosome diversity and mitochondrial DNA; contrasting the 'fast-train' (ETP model set) and the 'entangled-bank' (BAII set) models and finding evidence of deep ancestry of Polynesian Y-chromosomes in Near Oceania (Capelli *et al.*. 2001, Hurles *et al.*. 2002, Kayser *et al.*. 2001, Kayser *et al.*. 2000). Recent historic-period contributions to the Y-chromosome gene pool in Polynesia have also been detected with sources suggested in Native America and Europe (Hurles *et al.*. 1998, Hurles *et al.*. 2003).

One study found no evidence for shared haplotypes between Polynesians and Taiwanese, although all of the haplotypes in Taiwan and Oceania could be found in Southeast Asian populations (Su *et al.*. 2000), and the authors proposed a model of independent migrations from Southeast Asia to found populations in Taiwan and the Polynesian ancestral population after migration through Island Southeast Asia. Capelli *et al.* (2001) also concluded that many of the haplotypes found in populations from Southeast Asia, Oceania, southern China and Taiwan had an origin in island Southeast Asia and Melanesia and suggested that the dispersal of the Austronesian languages was mainly a cultural process, and that the Austronesian-speaking peoples have a Pleistocene-era paternal ancestry in Southeast Asia.

A phylogeny for human Y-chromosome variation continues to be developed as new markers are discovered, and the main branches emerging from studies to date have been labelled alphabetically from A to R (Jobling

Population	n	% C	% F/K/M	% F/K/NO	% F/K/P	% Other
Remote Oceania	892	34	14	19	4	30
Near Oceania	476	11	54	10	0	24
Island SEA	387	11	2	68	5	14
Taiwan & Asia	179	4	0	94	1	1
Australia	95	65	0	1	6	27

Table 1.1 Distribution of major Y-chromosome haplogroups in Oceania and surrounding regions

and Tyler-Smith 2003). A basic outline of this tree is shown in Appendix E1.2, which tabulates the combined results from five recent analyses of Y-chromosome diversity involving Pacific populations (Cox and Lahr 2005, Fris 2006, Hurles *et al.* 2005, Kayser *et al.* 2006, Underhill *et al.* 2001). The different markers used by each of these analyses to determine haplotypes limits the resolution of the combined results; however a broad summary of the haplotypes (Table 1.1) shows that haplogroups C and F/K/NO which are common within Oceania are also seen in neighbouring populations, while the F/K/M haplotypes common in Oceania are rare outside of the region. Kayser *et al.* (2006) reported Y-chromosome haplotype data for 1348 samples from populations in Oceania, Australia, Island and Mainland Southeast Asia, East Asia. As the authors suggested the origins of the C subhaplogroup lie in Oceania, the F/K/M haplogroup in Near Oceania, and the common F/K/NO/O subhaplogroup within Asia, they concluded that 65.8% of Polynesian Y-chromosomes could be traced back to Melanesia, and 28.3% to Asia. The distribution of the F/K/M haplogroup in particular (Table 1.1) fits well with an hypothesis of Near Oceanic origin, and if this continues to be supported as more markers are developed it indicates a substantial contribution to the existing Remote Oceanic populations from Near Oceanic ancestors.

Mitochondrial DNA analyses

Mitochondrial DNA (mtDNA) studies of Oceanic populations have found that Polynesian individuals have low sequence diversity, with one haplotype in the non-coding first hypervariable region (HVR-I) of the mtDNA becoming known as the 'Polynesian motif' (PM) due to its high frequencies in Polynesia (Ballinger *et al.* 1992; Hagelberg and Clegg 1993, Hagelberg *et al.*. 1994, Lum *et al.*. 1994, Melton *et al.*. 1995, Redd *et al.*. 1995, Sykes *et al.*. 1995, Murray-McIntosh *et al.*. 1998, Richards *et al.*. 1998, Hagelberg *et al.*. 1999b). The motif consists of four distinctive changes from the human mtDNA reference sequence, and is seen in combination with a 9 base-pair deletion in a non-coding region between COII and tRNA genes (Hertzberg *et al.*. 1989).

Several studies have traced variants of this motif in Near Oceania, Island Southeast Asian and Asian

populations, using the distribution of the motif and its immediate precursors to test support for and against the ETP and BAII model sets (Redd *et al.*. 1995, Sykes *et al.*. 1995, Lum *et al.*. 1998, Melton *et al.* 1995, Hagelberg *et al.*. 1999b). Other mtDNA haplotypes found at lower frequencies in Remote Oceania are also present in Near Oceania, where a greater diversity of sequence types is seen. A supplementary table (Appendix E1.3) summarises details of 32 studies of human mtDNA from Oceanic populations published between 1989 and 2006.

With developments in sequencing technology an increasing number of entire mtDNA sequences have become available in recent years. Entire mtDNA sequences from Australia and Oceania have confirmed the patterns of considerable divergence between these two regions indicated by HVR-I sequences (Ingman and Gyllensten 2003, van Holst Pellekaan *et al.* 2006), and clarified branching patterns in lineages, P and Q which are common in Near Oceania. Taiwanese samples with the pre-Polynesian Motif have demonstrated a close relationship between these haplotypes and Polynesian motif haplotypes from Polynesia (Ingman and Gyllensten 2003, Trejaut *et al.* 2005). Other entire mtDNA sequences from Near Oceania have revealed a number of deep lineages which are found only in this part of the world, reflecting the complexity expected of a region with such a long settlement history (Friedlander *et al.* 2005, Friedlaender *et al.* 2007, Merriwether *et al.* 2005).

1.4 Thesis outline

This thesis examines the evidence of Pacific prehistory retained in contemporary mitochondrial DNA variation through the sequencing of entire mitochondrial (mt) genomes from samples from Oceania and Taiwan. Chapter Two describes the collection of sequence information from 20 individuals for this project, and the phylogenetic analysis of the sequences as part of a larger mtDNA coding-region data set comprising all available mt genomes from Oceania, Australia, Island Southeast Asia and Taiwan. The computational demands for analysing such a large data set in terms of both sequence length and number of taxa are such that an exhaustive search of tree space is not feasible and an heuristic approach is required. The MinMax Squeeze (MMS) parsimony analysis (Holland *et al.* 2005b) used for the Oceanic data sets (Pierson *et al.* 2006, updated here to include recently described sequences from Australia, van Holst Pellekaan *et al.* 2006) enabled heuristic parsimony search scores to be evaluated according to an optimisation criterion.

In Chapter Three the haplogroups from Oceanic populations identified from the consensus networks generated in Chapter Two are analysed in greater detail, with phylogenetic analyses of the entire sequences including the control region. The haplogroups are divided to two groups according to their distribution:

'ancient' haplogroups with divergent haplotypes that are found only in Oceania, and have most recent common ancestry with other haplogroups at the level of macrohaplogroups M and N/R; and 'young' haplogroups which are closely-related within Oceania, and members of which are also found in neighbouring regions. Dates have been estimated for several ancestral vertices in the mtDNA phylogenies, and the implications and reliability of these are discussed.

The analyses of the Oceanic mt genome sequences revealed many instances of homoplasy; mutations occurring at the same base in parallel in different lineages, or 'multiple hits', where the same base has undergone changes repeatedly along a single path. These patterns were particularly evident in the fast-evolving control region, where several bases have been previously identified as 'hypervariable' (Stoneking 2000, Meyer *et al.* 1999). While the mt genome sequences provide high resolution phylogenetic information the overall sample size is small, and they are generally targeted for sequencing based on prior knowledge of control region variation and therefore not a random sample of the population from which they are derived. Large numbers of HVR-I sequences from Oceania are available from public databases and it was an aim for this project to review these sequences in light of the information gained from the entire mt genome phylogenies.

Chapters Four and Five describe the use of a large data set (1736 haplotypes) of human mt genomes from public data bases to explore the extent of homoplasy in a global sense, and its potential causes including the effects of selection and recombination events. Chapter Four reviews features of mitochondrial DNA, its function, replication and mode of inheritance, and describes the collation of the data set and variation within it. Finally, the results of tests of selection and recombination within subsets of the data are presented.

In Chapter Five a phylogenetic approach is taken to examine the occurrence of repeat mutations in mtDNA, focusing on the control region. Over a hundred random subsets of 75 taxa from the global data set were used to generate sets of parsimony trees from the coding region, which were tested for optimality using the MMS. The control region was mapped onto tree sets where the heuristic search score was proved optimal and the number of steps required to fit the trees for each base assessed. This results from this analysis were then used to devise weighting schemes for the HVR-I which were assessed in a second phylogenetic analysis. The relative performance of the coding and control regions (weighted and unweighted) at distinguishing known haplogroups was tested using 5000 data sets of fifteen taxa each containing five sequences randomly selected from three of 18 defined haplogroups.

Chapter Six returns to analyses of Oceanic mtDNA sequences, with the description of control region

sequences obtained from a sample set of 46 Polynesian individuals from Auckland, New Zealand. The mtDNA of all but one of these samples belonged to haplogroup N/R/B4a, the most common type found throughout Polynesia. Single nucleotide polymorphisms (SNPs) of interest in the coding region for further study of these sequences were identified from the B4a mt genome phylogeny reconstructed in Chapter Three, and it was possible to assess for the first time how closely related pre-Polynesian motif sequences from Polynesia are to the Polynesian motif sequences. The Auckland sequences were incorporated into two HVR-I data sets of different lengths: the first comprising >4000 HVR-I sequences gathered from public databases from Oceanic, Asian and American populations, and the second almost 1200 shorter sequences from Oceania alone. Haplotype distributions and phylogenies for the three main haplotypes found in Oceanic populations are presented.

Finally, Chapter Seven summarises the findings from the previous chapters, and discusses areas for further research which have been identified over the course of this project.

2. Analysis of the Oceanic data set

This chapter describes the collection and sequencing of mitochondrial (mt) genome samples from Oceania and Taiwan and their phylogenetic analysis within a larger data set containing sequences from Oceania, Taiwan, Island Southeast Asia and Australia (the 'Oceanic' data set). The sequences from this study include the first mt genomes reported from Vanuatu and Micronesia, and have been analysed using a novel phylogenetic approach, the Min Max Squeeze (MMS, Holland *et al.* 2005). This method is suited to population-level data, as it allows large data sets such as this to be analysed relatively quickly with an optimality criterion. This analysis has been published (Pierson *et al.* 2006, Appendix A) and here the initial study is summarised and updated to include revised sequences from the Andaman Islands and new mt sequences from Australia. In the following chapter the haplogroup subsets defined by the Oceanic analysis are explored in greater detail.

2.1 Collection of new sequences

Sample sources

This project extends initial work begun at the Allan Wilson Centre for Molecular Ecology and Evolution (AWCMEE) at Massey University in Palmerston North in 2002. Postdoctoral researchers Matthew Hurles and Rosa Martinez-Arias, working with Professor David Penny identified a set of mtDNA HVR-I haplotypes found in Remote Oceanic populations from the literature, and sourced samples to represent these types from Professor John Clegg's collection at the Weatherall Institute of Molecular Medicine, University of Oxford and from Matthew Hurles' samples.

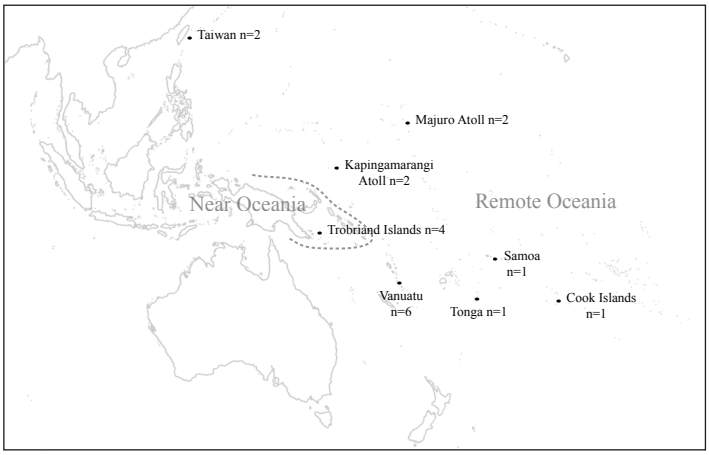


Figure 2.1 Sample locations

The most common HVR-I haplotype found in Polynesian samples has become known as the 'Polynesian motif', and has a distinctive set of transitions relative to the revised Cambridge reference sequence (rCRS; Andrews *et al.* 1999): 16217C, 16247G and 16261T. As the initial intention was to sequence just ten mt genomes, only one sample (working code TL36) with the Polynesian motif was selected from the University of Oxford collection.

Accession number	Working code	Geographic origin	Haplogroup
DQ372868	AMI15	Taiwan[1]	M/M7c
DQ372869	PAI9	Taiwan[1]	N/R/B5a
DQ372870	TRO122	Trobriand Islands, Papua New Guinea[2]	N/R/P2
DQ372871	TRO131	Trobriand Islands, Papua New Guinea[2]	N/R/B4a1a
DQ372872	TRO133	Trobriand Islands, Papua New Guinea[2]	N/R/P2
DQ372873	TRO137	Trobriand Islands, Papua New Guinea[2]	N/R/B4a1a1
DQ372874	KAP19	Kapingamarangi Atoll, Caroline Islands[1]	N/R/B4a1a1
DQ372875	KAP89	Kapingamarangi Atoll, Caroline Islands[1]	N/R/B4a1a1
DQ372876	MJ22	Majuro Atoll, Marshall Islands[1]	M/M7c
DQ372877	MJ86	Majuro Atoll, Marshall Islands[1]	N/R/B4a1a1
DQ372878	PO314	Espiritu Santo, Vanuatu[3]	N/R/B4a1a1/Pol. motif
DQ372879	PO332	Espiritu Santo, Vanuatu[3]	M/M28
DQ372880	PO392	Espiritu Santo, Vanuatu[3]	M/Q1
DQ372881	MF025	Maewo, Vanuatu[3]	N/R/B4a1a1/Pol. motif
DQ372882	MO304	Vanuatu[1]	M/Q1
DQ372883	T726	Vanuatu[1]	M/M28
DQ372884	CI153	Cook Islands[3]	M/Q1
DQ372885	WS72	Samoa[1]	M/Q1
DQ372886	TL36	Tonga[1]	N/R/B4a1a1/Pol. motif
DQ372887	TRI65	New Zealand (European)[4]	N/W

Table 2.1 New sample details
[1] Supplied by J.B. Clegg, Weatherall Institute of Molecular Medicine, University of Oxford
[2] Provided by W. Schiefenhovel
[3] Samples from Matthew Hurles
[4] Methodological control sample, provided by P.A. McLenachan

The remaining nine samples represented the diversity found in HVR-I sequences and came from Taiwan, Micronesia, Vanuatu and Polynesia. In total twenty mt genomes were sequenced over the course of this study (Table 2.1). Additional samples from Vanuatu and Cook Islands were provided by Matthew Hurles (Wellcome Trust Sanger Institute, Cambridge, UK) and four samples from the Trobriand Islands, off the eastern coast of Papua New Guinea were sourced from Dr. W. Schiefenhovel (Human Ethology, Max-Planck-Institute, Andechs, Germany).

Table 2.1 lists the working codes, accession numbers and sources of the samples sequenced. Thirteen of the twenty mt genomes are from Remote Oceania, four from Near Oceania, two from Taiwan and one from a New Zealander of European ancestry. The locations of the samples are shown in Figure 2.1.

Amplification, sequencing and assembly of whole mt genomes

Eight of the samples (CI153, MF025, PO314, PO332, PO392, TRI65, TRO122 and TRO131), were sequenced at the AWCMEE by Rosa Martinez-Arias in 2002-2003, with some further re-sequencing and

editing required by myself at the University of Canterbury where the remaining 12 samples were sequenced. Briefly, the methodology involved an initial polymerase chain reaction (PCR) amplification of the entire mt genome in two large overlapping fragments of ~10.8kb and ~7.5kb, using combinations of primers from a set of 24 designed to amplify the entire mtDNA genome by (Rieder *et al.* 1998). The long PCR is a safe-guard against the unintentional amplification of nuclear copies of mtDNA (NuMTs; Bensasson *et al.* 2001, Parr *et al.* 2006), and was also important, particularly in the case of the samples from the University of Oxford collection, in allowing the several subsequent PCR reactions to be carried out from a small amount of sample template. From the long PCR products internal fragments of ~2kb were amplified, and direct sequencing carried out on these using both the PCR and internal primers. See Appendix C2.1 for details of the PCR reactions and the primers used (primers are also marked on the annotated reference sequence, Appendix B).

Chromatograms were edited and sequences assembled using Sequencher™ (Version 4.2.2, Gene Codes Corporation). The number of sequencing reads covering each base position varied for the different samples; the set of primers is designed to overlap when long sequences (~750-1000 bases) are obtained and in the majority of cases each base position was covered by at least two reads, from forward and reverse directions. Figure 2.2 shows an example of one of the assembled contigs, for sample PAI9 (DQ372869). Exceptions to this coverage are two areas in the control region (nt16183-nt16192, nt303-nt310) which in some lineages are uninterrupted poly-cytosine sequences. It was often only possible to get sequence information from a single direction on either side of these, as the sequence reads failed once into the cytosine repeat regions.

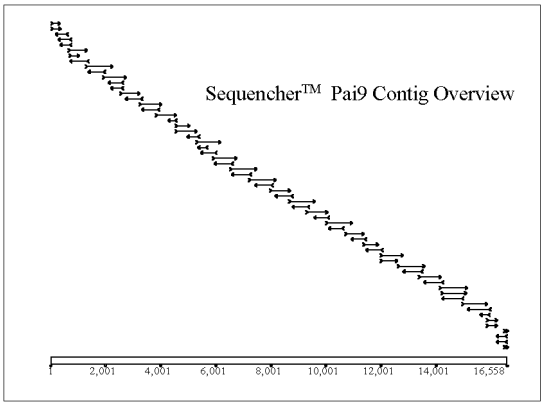


Figure 2.2 Contig example PAI9 (DQ372869)
This overview shows the sequence coverage in forward and reverse directions (62 sequences) of the mt genome sequence DQ372869. The base numbering is according to the rCRS (AC_000021.2).

In recent years several reports have drawn attention to potential and actual errors in human mtDNA data due both to automated sequencing mistakes in base assignments and human error in downstream data manipulation; see for example: Bandelt *et al.* 2002, Bandelt *et al.* 2003, Bandelt *et al.* 2004a, Bandelt *et al.* 2004b, Bandelt *et al.* 2005, Bandelt and Kivisild 2006, Bandelt *et al.* 2007, Forster 2003, Yao *et al.* 2003. The mt genomes generated in this study were carefully checked, with each chromatogram edited manually. Summary lists of polymorphisms relative to the rCRS for the

twenty sequences from this study are provided in Appendix E2.1. There were some unexpected variants in four samples: for example DQ372868 (AMI15), from Taiwan, belongs within haplogroup M/M7c but has the 9bp deletion which is characteristic of haplogroup N/R/B. These unusual polymorphisms were verified by repeat sequencing, and are described in Appendix E following the sequence polymorphism lists.

2.3 Data set details and phylogenetic methods

Data set details

The 20 sequences generated from this project were manually aligned using SE-AL (Rambaut 1996) with others available on public databases, as part of a large data set assembled between 2003-2006 (details of this alignment follow in Chapter 4). From this data set a subset of 137 geographically relevant sequences were selected for phylogenetic analysis.

This Oceanic data set contained an African L3a sequence, (AF347014, Ingman *et al.* 2000); and all sequences from Taiwan, Island Southeast Asia, Oceania and Australia. Eight additional sequences from Australia (van Holst Pellekaan *et al.* 2006) published recently have been incorporated into a second revised Oceanic data set, bringing the total number of sequences to 145. Several sequences from the Andaman Islands have been amended since the first Oceanic analysis, and the corrected versions have replaced the originals in the new data set. Table 2.2 summarises the geographic origin of the samples in the first and second Oceanic data sets. The two data sets are distinguished by their number of unique coding-region haplotypes (excluding the control region nt16024-nt576); the initial data set of 137 individuals contained 127 coding-region haplotypes while the revised version has 133.

Maximum parsimony, the MinMax Squeeze and consensus networks

PAUP* (version 4.0b10, Swofford 2003) was used to find the most parsimonious trees for the Oceanic data sets by heuristic search, after excluding any gapped characters (branch swapping = Tree Bisection-Reconnection, stepwise addition = simple). As there is known to be a distinctive nine base-pair deletion of one copy of a tandem repeat in an intergenic region at nt8270-nt8294 in individuals with the 'Polynesian motif' this was further encoded in the data set by adding a transition where it occurred. It is not computationally feasible to assess all possible trees through an exact search - for number of taxa=n, there are (2n -5)!! possible binary trees (Semple and Steel 2003); here for 127 taxa this is equal to ~4 x 10^{245} trees - and the heuristic approach can not guarantee to return all of the most parsimonious trees; however the parsimony score of the set of trees returned can be evaluated using the MinMax Squeeze programme (Holland *et al.* 2005b).

Location	n	Accession numbers	Source
Australia	26 (34)	AF346963-AF346965	Ingman et al 2000
		AY289051-AY289067	Ingman & Gyllensten 2003
		DQ112750-DQ112755	Kivisild et al 2006
		DQ404440-DQ404447 (note: vers.3)	Van Holst Pellekaan et al 2006
Island Southeast Asia			
Philippines	2	AF382012	Maca-Meyer et al 2001
		AY289070	Ingman & Gyllensten 2003
Malaysia	9	AY963576-AY963584	Macaulay et al 2005
Nicobar Islands	5	AY950286-AY950290	Thangaraj et al 2005
Andaman Islands	10	*AY950291-AY950300*	Thangaraj et al 2005
Taiwan	14	AY289095-AY289098	Ingman & Gyllensten 2003
		AJ842744-AJ842751	Trejaut et al 2005
		DQ372868-DQ372869	present study
Near Oceania			
New Guinea	25	AF347002-AF347005	Ingman et al 2000
		AY289076-AY289092	Ingman & Gyllensten 2003
		DQ112895-DQ112898	Kivisild et al 2006
Bismarck Archipelago	14	AY956412-AY956414	Friedlaender et al 2005
		DQ137398-DQ137404, DQ137406-DQ137409	Merriwether et al 2005
Bougainville	5	AY289075	Ingman & Gyllensten 2003
		AY963574	Macaulay et al 2005
		DQ137405, DQ137410-DQ137411	Merriwether et al 2005
'Melanesian'	3	DQ112885-DQ112887	Kivisild et al 2006
Trobriand Islands	4	DQ372870-DQ372873	present study
Remote Oceania			
Samoa	4	AF347007	Ingman et al 2000
		AY289093-AY289094	Ingman & Gyllensten 2003
		DQ372885	present study
Tonga	2	AY289102	Ingman & Gyllensten 2003
		DQ372886	present study
Cook Islands	3	AY289068-AY289069	Ingman & Gyllensten 2003
		DQ372884	present study
Vanuatu	6	DQ372878-DQ372883	present study
Kapingamarangi Atoll	2	DQ372874-DQ372875	present study
Marshall Islands	2	DQ372876-DQ37287	present study
Total	136 (144)		

Table 2.2 Oceanic data set sequence details.
The new Australian sequences included in the revised Oceanic data set (Oceanic-133), and the sequences from the Andaman Islands revised in mid-2006, are shown in italics. The bracketed numbers reflect the increase in number of sequences from the original Oceanic data set (Oceanic-127) to the revised one (Oceanic-133). Twenty-two sequences were reduced to ten haplotypes across the coding region in the Oceanic-133 data set. These were: AJ842749 identical to DQ372871, AY289077 identical to AY289102, AY950286 and AY950287 identical to AY950288, AY950291 and AY950292 identical to AY950295, AY950298 identical to AY950300, DQ137402 identical to DQ137404, DQ137408 identical to DQ137409, DQ372884 identical to DQ372885, DQ372874 identical to DQ372875, DQ112885 identical to DQ112887.

The MinMax Squeeze takes the parsimony score found by the heuristic search as an upper bound and derives a lower bound by summing the parsimony scores of partitions of the data set. If the upper and lower bounds meet then the most parsimonious trees found by heuristic search are proved optimal.

Consensus networks (Holland and Moulton 2003; Holland *et al.* 2005a) combine sets of trees within a single graph, allowing areas of disagreement to be highlighted. Where two different branching possibilities exist both are displayed in the consensus graph by representing each as a pair of parallel edges. The complexity of the area of conflict increases with the number of different branching possibilities, forming high dimensional hypercubes. Consensus networks were constructed using a Python script written by Barbara Holland to convert PHYLIP format trees to a NEXUS format file of splits, and then drawn using Spectronet 1.27 (Huber *et al.* 2002). SplitsTree 4 (Huson and Bryant 2006) also implements the consensus network algorithm, and will read NEXUS format trees. All splits (edges) in the sets of equally parsimonious trees are shown in the consensus networks.

2.4 Results

Consensus networks for the original Oceanic-127 and the revised Oceanic-133 data sets are presented in Figures 2.3 and 2.4. The changes to the data set make a considerable difference to the number of trees found by heuristic search, but only minor differences to the branching structure of the phylogeny overall.

Oceanic-127

The heuristic search for most parsimonious trees found 582 624 trees with score 412 for the coding-region Oceanic-127 data set (parsimony informative characters = 282), and these are displayed in the consensus network in Figure 2.3a. Apart from two areas of conflict at the M and N/R vertices the network is largely tree-like, with the only other reticulation occurring in the M/Q haplogroup. It is evident that most of the trees found differ in the branching orders from the deep ancestral M and N/R vertices, both of which are known to have many descendant lineages (for example see Appendix Figures D3.5 and D3.6 which show M and N haplogroups found in East Asia).

The lower bound reached by MMS for Oceanic-127 was 410. As high levels of homoplasy have been shown to affect the efficiency of the MMS algorithm in finding the maximum partition score (Holland *et al.* 2005b) the number of steps required on the trees for the parsimony informative characters was investigated: a score of 1 indicates the character has two states in the data set, and only requires one change between these. The average number of steps required for each character over all trees found was calculated using

PAUP* and Microsoft Excel®. Five of the 282 characters were highly homoplasious; requiring 5 or more steps on average across the trees: nt709 (6.8), nt1598 (6.2), nt1719 (5), nt10398 (5) and nt15924 (5). In several sequences within the M lineages present in the data set (M27, M28a, M29 and M42) there is a transition from G to A in the 12S rRNA gene at nt1598, which in context with other base changes appears to be recurrent rather than ancestral to these individuals. A similar pattern is found at the N/R vertex, where several lineages descending from the N/R vertex appear to have a back mutation at nt10398 from A to G (as a non-synonymous transition at nt10398 from G to A in the ND3 gene is generally represented as one of 5 substitutions that define the N macrohaplogroup, Appendix D3.5).

When the two sites showing recurrent mutations described above, nt1598 and nt10398, were excluded from the parsimony analysis for Oceanic-127, 165 trees were found by heuristic search (parsimony informative characters=280, search score=399) and this upper bound of 399 was met by the MMS program, guaranteeing the parsimony score optimal for the reduced data set. The consensus network of the 165 most parsimonious trees is shown in Figure 2.3b. Excluding nt1598 and nt10398 removes all conflict between the trees at the M vertex, and greatly reduces the branching possibilities at N/R. The area of conflict at N/R involves single individuals representing the recently described R12 and R21 haplogroups (Kivisild *et al.* 2006, Macaulay *et al.* 2005), an Australian N/R/P sequence and the remainder of the N/R/P haplogroup. Several base changes are responsible: there are two shared coding-region transitions between the R21 and P7 sequences at nt12361 and nt15613, a substitution at nt11404 is found in both the R12 and R21 sequences, while the P-defining substitution at nt15607 is found in all P sequences but not in the R12 and R21 sequences. Features of the P haplogroup and the relationships between the P, R12 and R21 sequences are described in more detail in Chapter 3.

Oceanic-133

The increase in number of haplotypes by six from the Oceanic-127 to the Oceanic-133 data set and the inclusion of the revised Andamanese sequences increased the number of parsimony informative characters from 282 to 299. The initial heuristic search on the Oceanic-127 data set excluding the control region found 582624 most parsimonious trees, and provided an upper bound for the MMS of 412. The search on the Oceanic-133 data set reached the maximum number of trees (1 761 200) able to be stored on the computer used (Intel Pentium® 4CPU 3.2GHz processor, 2.87GB RAM, time taken 132.5 hours), with a parsimony score of 452, and the maximum lower bound reached by MMS was 448.

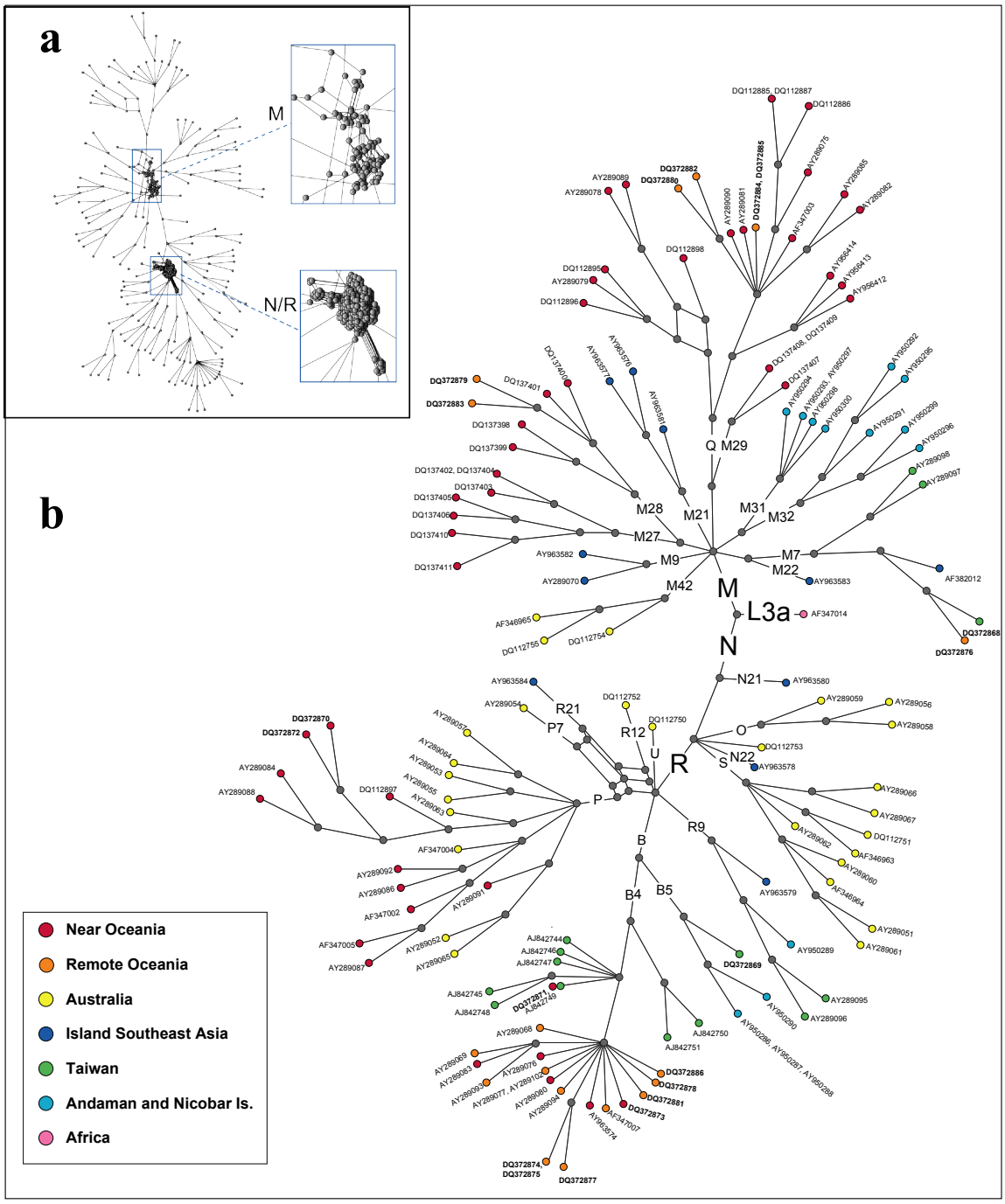


Figure 2.3 Oceanic-127 consensus networks
a) Consensus of 582,624 most parsimonious trees found by heuristic search; upper bound 412, lower bound 410. The entire coding region (282 parsimony informative characters) of the mtDNA sequence of 127 haplotypes was analysed. The two major areas of conflict amongst the trees at the M and N/R vertices are enlarged. b) Consensus network of 165 guaranteed optimal most parsimonious trees found by heuristic search when two characters, nt1598 and nt10398, were excluded from the analysis. The MinMax Squeeze approach guarantees the heuristic search score of 399 to be minimal. Accession names of sequences determined from this project are shown in bold; haplogroups are labelled according to existing nomenclature.

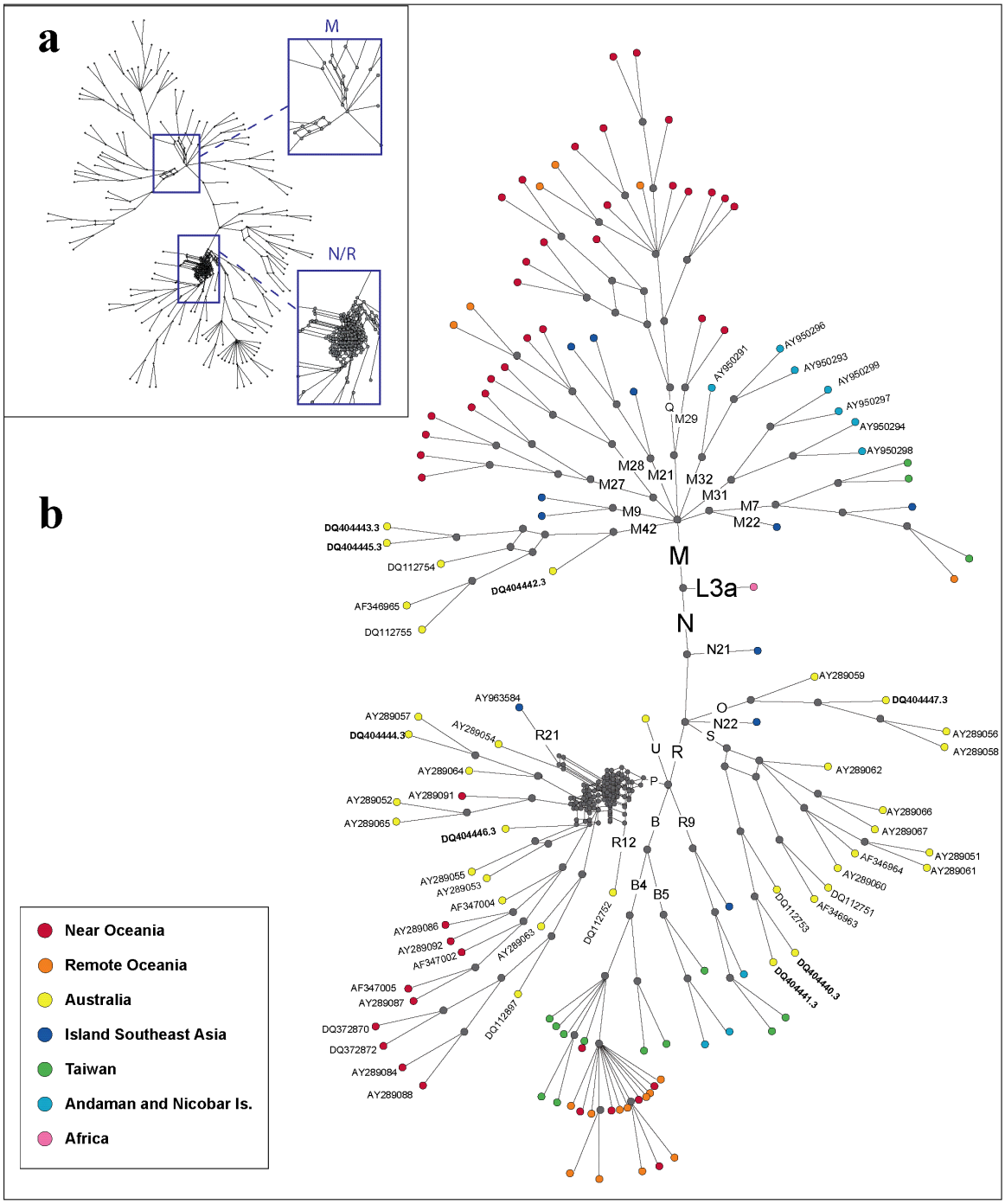


Figure 2.4 Oceania-133 dataset consensus networks
a) Consensus of 1 761 200 most parsimonious trees found by heuristic search (maximum trees limit reached); upper bound 452, lower bound 448. The entire coding region (299 parsimony informative characters) of the mtDNA sequence of 133 haplotypes was analysed. The two areas of conflict at the M and N/R vertices are enlarged for contrast with Figure 2.3a: while the branching possibilities are reduced at the M vertex, those at N/R have increased. b) Consensus network of 9982 most parsimonious trees found by heuristic search (parsimony score=438) when two characters, nt1598 and nt10398, were excluded from the analysis. The lower bound reached by the MinMax Squeeze was 437. The new Australian sequences (van Holst Pellekaan et al 2006) are shown in bold and only sequences within haplogroups showing changes to the Oceanic-127 consensus network are labelled.

The consensus network of the Oceanic-133 trees from the entire coding region (Figure 2.4a) is similar to the networks for the Oceanic-127 data set (Figure 2.3). Most haplogroups have the same topology, with conflicts appearing at the deeper N/R and M vertices; but the revisions to the sequences from the Andaman Islands appear to have made a significant difference to the complexity of the conflict in the M haplogroup. The two haplogroups found in the Andamanese, M31 and M32 (Thangaraj *et al.* 2005) now branch directly from the M vertex, with conflicts arising in the branching from M of haplogroups Q, M21 and M9. Haplogroups M27, M28, M29 and M42 branch together from M, with uncertainty in the subsequent branching order. Three of the new Australian sequences fall into the M42 haplogroup, and there is also a single area of conflict between trees within this haplogroup.

It appears that most of the trees found differ in the branching from N/R of its descendant lineages; particularly, as was also the case for Oceanic-127, in the structure of relationships between the P, R21 and R12 haplogroup sequences. The five new Australian sequences belong to three haplogroups: N/O, N/S and N/R/P, and the two sequences within N/S introduce two areas of uncertainty within this haplogroup.

As the lower bound found by the MMS was 4 steps off the upper bound for the entire coding region an heuristic search was run excluding the characters nt1598 and nt10398 which had proved to be highly homoplasious in the Oceanic-127 data set. Now with 297 parsimony informative characters the search resulted in only 9982 trees being found, with a score of 438. The MMS reached a lower bound of 437 for this data set. The consensus network (Figure 2.4b) is less complex with these two characters excluded, as was the case for the original data set.

2.5 Discussion

Geographic distribution of haplogroups

There is marked geographic clustering within the Oceanic consensus networks. The N/R/P haplogroup contains haplotypes from Australia and Near Oceania, haplogroups M/M42, N/O and N/S are currently found only in Australia and M/Q, M/M29, M/M27 and M/M28 are present only in Oceania. The eight new Australian sequences in Oceania-133 fall within all the haplogroups found in Australia. Other geographically restricted haplogroups are M/M31 and M/M32, found in samples from the Andaman Islands and M/M22, M/M21, N/N21, N/N22 and N/R/R21 found only to date in Malaysian aboriginal populations. By contrast, the N/R/B4, N/R/B5, N/R/R9, N/R/U, M/M7 and M/M9 haplotypes are widespread within the area covered and are also present in other global populations. Their distribution suggests they may represent more recent migrations, of ‘young’ haplogroups, into areas initially settled by the ancestors of those carrying the

geographically restricted haplotypes.

Monophyly of previously described M haplogroups

Four of the seven branches from the M vertex in the consensus network for Oceanic-127 combine haplogroups that have been previously reported as direct descendants of M. M31 and M32 are found to date only from individuals from the Andaman Islands, and as reported in Pierson *et al.* (2006) a single transition, from A to G at base nt1524 in the 12S rRNA gene linked the two haplogroups which had been initially described as branching directly from the M vertex polytomy (Thangaraj *et al.* 2005). When the revised sequences were analysed in Oceanic-133 both haplogroups branch from the M vertex, and the internal branching patterns also differ; M31 becomes more resolved, while the M32 haplotypes collapse from five to three (comparisons are between Oceanic-127 and Oceanic-133 networks, without nt1598 and nt10398).

Haplogroups Q and M29 are linked by a branch to the M vertex as they share a synonymous transition in the ND5 gene at nt13500. Merriwether *et al.* (2005) have suggested that more sequences, particularly from M29, are required to assess whether this polymorphism reflects shared ancestry or has been acquired independently in both Q and M29. A transition at nt13500 occurs in parallel in several other global lineages. Similarly, the basal transition linking M27 and M28 (at nt1719 in the 16S rRNA gene) has arisen several times independently (for example in L3e, M/M8/C, M/D, N/R/B4b, N/R/P) and does not provide strong support for an ancestral link between M27 and M28.

The fourth grouping of two M haplogroups previously described independently is between the M7 sequences and M22, represented by a single sequence from Malaysia. These are grouped by two shared polymorphisms: a synonymous transition from A to G at nt5351 in the ND2 gene and a non-synonymous transition from A to G at nt15236 in the Cyt. *b* gene. Both substitutions are present in the M22 sequence, while the nt5231 transition is found in the M7b Taiwanese sequences, and the nt15236 polymorphism in the Micronesian and Taiwanese M7c sequences. In the following chapter this association is examined in greater detail in an entire mtDNA (coding and control) analysis of the M22 and M7 sequences from the Pacific, with other M7 sequences from mainland Asia.

The MMS analysis

A data set of this size, in terms of both number of sequences and number of characters, presents considerable computational challenges for phylogenetic reconstruction. Within the relatively shallow timeframes of intraspecific analyses mtDNA sequences remain closely related and the phylogenies can be expected to differ significantly from interspecific reconstructions: ancestral sequences may still be extant within population

samples and true (hard) polytomies are common.

Distance-based methods have been applied in studies of similar size (Ingman and Gyllensten 2003, Ingman *et al.* 2000, Mishmar *et al.* 2003). While neighbor-joining (NJ) is computationally fast it provides only a relatively rough estimate. Maximum likelihood methods have also been used; but on smaller data sets of up to 31 sequences (Macaulay *et al.* 2005b, Torroni *et al.* 2001), as the computational demands for these analyses is very high. Bayesian likelihood approaches are better suited to larger data sets (Atkinson 2006) and may be used more in future.

Several analyses of entire mtDNA sequences (for example Finnila *et al.* 2001, Herrnstadt *et al.* 2002, Moilanen and Majamaa 2003) have used the median-joining, or reduced median network method first proposed for use with RFLP and HVR-I data (Bandelt *et al.* 1995, Bandelt *et al.* 1999, Bandelt *et al.* 2000). This network approach defines splits in the data set: each variable character divides the taxon set into two (this requires pre-processing of the DNA characters in the case of 3- or 4-state characters), and these splits are visualised in a graph. Where all the splits are compatible this graph will be a tree; when there are conflicting signals in the data the graph (network) becomes more complex. Unmodified median networks have been demonstrated to contain all of the most parsimonious trees in a data set (Bandelt *et al.* 1995), but without reduction techniques the resulting network is often difficult to interpret as it can contain high dimensional cubes due to uncertainties in branching order. The resolution of reticulations in the reduced median network is based on the frequency of individuals carrying each opposing split, and the number of characters contributing to the split (Bandelt *et al.* 1995).

The MMS method used in this study has the advantage over distance and network methods of having an optimality criterion, and is a rapid means of analysing large data sets compared to maximum likelihood methods. The disadvantage of using the parsimony method on intraspecific data sets noted by Clement *et al.* (2000: p1657): ‘it tends to generate a cumbersome amount of most parsimonious trees at the population level’ is overcome by the representation of these many trees in a consensus network. These networks, which are constructed from the same split-based algorithm of median networks but use a set of trees, rather than characters, as input, can provide a clear and concise summary of the information contained within the many trees highlighting areas of conflict and agreement. An advantage of the method over the use of median networks is that the data set does not need to be pre-processed to set characters to splits and the information held in three or four state characters is maintained.

The Oceanic data set analysis provides an overview of mtDNA histories in the region. In the following chapter each haplogroup present in Oceania is examined in greater detail, in haplogroup data sets incorporating the entire mtDNA sequence from both the coding and control regions.

3. mtDNA haplogroups in Oceania

This chapter continues the analysis of whole mt genomes from Oceania, looking in detail at each of the haplogroups present. In Chapter Two the MinMax Squeeze analysis of all sequences from Oceania, Australia, Island Southeast Asia and Taiwan provided an overview of the relationships between the different haplogroups and their geographic distributions. The haplogroups in Oceania were assigned to two broad sets: here termed 'ancient' and 'young' following Friedlaender *et al.* (2007). The first of these, 'ancient' contains the haplogroups that are found only within Oceania and Australia, and share most recent common ancestry with other global types at the 'Out-Of-Africa' L3 polytomy. In the second, 'young', set are the haplogroups present in Oceanic populations which are also found outside of the region, and whose phylogenies and distributions suggest later migration into Oceania.

Haplogroups descending from the M, N and N/R vertices identified in the Oceanic consensus network (Figure 2.4) are analysed separately below, with sequences belonging to the 'young' haplogroups from regions outside of the coverage of the data set used in Chapter Two also included. The detailed descriptions of nucleotide changes within the haplogroups are intended as a guide for future work, enabling a screening approach to target a particular area of the phylogeny for more detailed analysis.

Dating phylogenies is a common practice in studies of human mtDNA, and molecular dating has great potential to add to existing means of timing events in human prehistory. This chapter explores issues in molecular dating in the Oceanic context, and in human mtDNA in general, looking at TMRCA (time to the most recent common ancestor) date estimates calculated using the rho dating technique (Forster *et al.* 1996) and three different subsets of mtDNA variation. Four mtDNA data sets not related to Oceanic prehistory, from Australia, Africa, the Americas and Europe, are included here for comparative purposes.

3.1 Methods

Phylogenetic analysis

Ten data sets were analysed, seven including sequences generated from this project (files are provided in Appendix F3.1 in NEXUS format). Each subset included either the L3a sequence present in the Oceanic data sets (AF347014) or the rCRS (AC_000021.2) to orient the trees and consensus networks, with the exception of the L1c data set which included an L1b sequence (AF346986). As these data sets were considerably smaller (9-46 haplotypes) than the Oceanic data set the entire mtDNA sequence, coding and control, could be analysed using the MMS and consensus network approach described in Chapter Two.

The Oceanic consensus network (Figure 2.4) was used to define and identify closely related haplogroups. The M27 and M28 haplogroup sequences were analysed in a single data set as they share a branch from the M vertex, and M29 sequences were included with the Q haplogroup data set to explore the connection between these two haplogroups. The conflicting branching possibilities in the N/R/P haplogroup in the Oceanic consensus networks were examined by including the N/R21 sequence with this subset, and adding the coding-only N/R12 sequence to the labelled tree. Similarly, as the M/M7 and M/M22 sequences share a branch from the M vertex these were analysed together.

Additional sequences from Asia not included in the Oceanic data sets were incorporated into the haplogroup sets for M/M7bc, N/R/B5 and N/R/B4. The N/W haplogroup (at present known only from Eurasia) was analysed with all available entire W sequences, as was the N/S haplogroup (geographically restricted to Australia), and all sequences belonging to a sub-haplogroup of L1, L1bc/c (found in African populations and outside of Africa in descendants of historic period African immigrants). A coding-region only data set of all sequences belonging to the N/R/B4bd subhaplogroup B4b (which contains American sequences named B2 in earlier studies, for example in Kivisild *et al.* 2006) was analysed to examine the date estimates for the MRCA of American and Asian B sequences. All of the above sequences which were not present in the Oceanic data sets (Chapter Two) were identified to haplogroup as part of a large global data set described in Chapter Four.

The consensus networks of most parsimonious trees found (or in some cases, the single tree) were used to guide the reconstruction of a single base-labelled tree. Lists of base changes relative to the Mitomap rCRS were generated by Sequencher™ (Version 4.2.2, Gene Codes Corporation, Ann Arbor, MI) and the trees drawn using Adobe® Illustrator® CS2 (©1987-2005 Adobe Systems Incorporated). Amino acid substitution details were obtained using "MitoAnalyzer", available at http://www.cstl.nist.gov/biotech/strbase/mitoanalyzer.html (2000, National Institute of Standards and Technology: Gaithersburg, MD, USA), with additional re-checking using the tables provided in the mtDB database (http://www.genpat.uu.se/mtDB/, Ingman and Gyllensten 2006).

The Kivisild *et al.* (2006) sequences (accessions beginning DQ112) are coding-region only, and while these were not included in the analyses, they have been added to the base-labelled phylogenies. As gapped sites are excluded from the MMS analysis, extensions and deletions at the cytosine tract between nt309 and nt315 have not contributed to the analyses and are therefore not reconstructed on the base-labelled trees. Another cytosine tract in the control region, between nt16184 and nt16193 also has length variations: most result in the transversion substitution of the flanking 5' 'AAAA' sequence with cytosine residues, but extensions to

the tract beyond nt16193 do occur and result in gaps in the alignment. These were also not included in the analyses, and are not reconstructed on the labelled trees. There is a short (CA) microsatellite between nt514 and nt523; the rCRS has 5 copies of this motif while others in the data sets vary from 4 to 7 copies. These variations did not contribute to the MMS analysis, but are reconstructed on the labelled trees. In order to allow the nine base-pair deletion of bases nt8280-8288 to act as a character in the parsimony analyses it was artificially encoded in the alignments by adding a guanine at nt8280 where the deletion occurred to force a transition.

It should be stressed that while the consensus networks display the set of optimal most parsimonious trees found, the base-labelled trees provide a single interpretation of branching order taken from this set and are intended as a visual summary of the polymorphisms present within the data set rather than necessarily the 'best' phylogenetic reconstruction. All data were carefully checked and re-checked in each phylogeny, to minimise errors. In the visual representations of haplogroup variation the maximum amount of information possible is presented on a single page while maintaining print legibility. Electronic searching and enlarging viewing functions (using Adobe® Acrobat® programs) are very useful for close examination of the phylogenies, and an electronic appendix of these figures is provided (Appendix F3.2).

Molecular dating

TMRCA calculations were derived from the base-labelled phylogenies using a method described by Forster *et al.* (1996). This approach takes the average of the number of substitutions from an ancestral vertex along each path to its descendants (the rho statistic) with a measure of the variance calculated as described by Saillard *et al.* (2000). Three rates described in the literature were used to calculate the dates. The first is calculated solely from synonymous substitutions in the protein-coding genes, including substitutions within stop codons that do not affect the function (Kivisild *et al.* 2006). The second rate takes all coding region changes into account (Mishmar *et al.* 2003). Both of these rates are calibrated by comparison to chimpanzee sequences, with an estimate of the TMRCA of human and chimpanzee mtDNA of 6.5 million years (Goodman *et al.* 1998). The third rate is calculated from transitions over the first hypervariable region (HVR-I) of the control region, from nt16090 to nt16365. It is estimated from Native American Eskimo and Na-Dene sequences and calibrated with a date of expansion related to the end of the Younger Dryas glacial relapse approximately 11300 years ago (Forster *et al.* 1996).

3.2 Results and Discussion

Table 3.1 summarises the results of the MMS analyses. The parsimony scores of all ten data sets were proved

Table 3.1 Haplogroup data set details

Data set	No. taxa	Pars. inform. chars.	Pars. score	MMS max.	No. trees
M/QwithM29	20	78	92	92	2
N/R/PwithR21	23	74	110	110	9
M27withM28	14	82	95	95	1
N/R/B4a	47	57	87	87	1274
N/B5a	11	17	21	21	1
M7bcM22	46	55	89	89	5040
L1bc/c	16	94	119	119	1
N/W	33	24	36	36	15
N/S	11	41	52	52	9
N/R/B4bd/b	24	16	18	18	3

minimal. The number of sequences within each data set ranged from 11 to 47, and in most instances the number of most parsimonious trees found was low; the notable exception being the M/M7bc with M/M22 and N/R/B4a data sets for which 1274 and 5040 most parsimonious trees were found respectively.

Each Oceanic haplogroup subset is discussed below in detail, beginning with the ancient Oceanic haplogroups M/Q with M/M29, N/R/P, and M/M27 and M28. The description of features of the young Oceanic N/R/B4a and M/M7bc haplogroups, and N/R/B5a, which contains a sample from Taiwan sequenced for this study follows. Finally, TMRCA estimates have been calculated for several vertices in the phylogenies and these are presented and discussed along with estimates from four other global haplogroups.

'Ancient' Oceanic haplogroups

M/Q and M/M29 haplogroups

As the M/Q and M/M29 haplogroups share a common branch from the M vertex in the Oceanic coding-region analysis (Chapter Two) these sequences were combined in an entire mtDNA analysis. Five M/Q coding-only sequences from Kivisild *et al.* (2006) were excluded, reducing the number of sequences in this data set to 19 (effectively 18 as two of the M29 sequences, DQ137408 and DQ137409 differ only in length

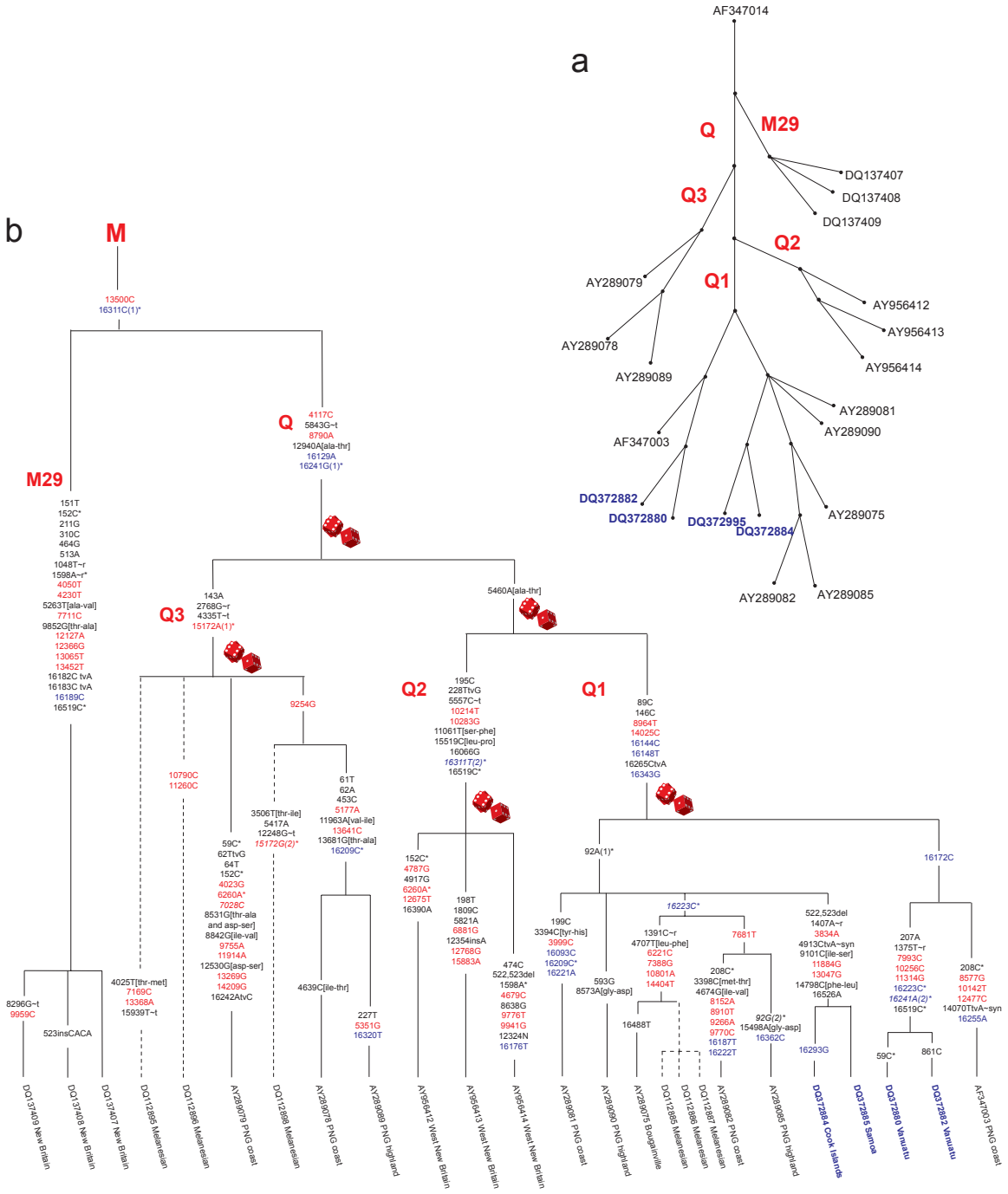


Figure 3.1 Q haplogroup with M29 consensus network and branch-labelled phylogeny
a) The consensus network of 2 most parsimonious trees, from 16 Q and 3 M29 individuals with AF347014 L3 outgroup found by heuristic PAUP* search, (version 4. 0b10, Swofford 2003), constructed using Spectronet (version 1.25 Huber et al 2002). The entire mtDNA sequence was analysed: there are 78 parsimony informative characters, and the parsimony score of 92 was proved minimal using MMS. Sequences from this study are shown in blue type.

Figure caption continues on following page

Figure 3.1 Q haplogroup with M29 consensus network (cont. from previous page)
b) A branch-labelled phylogeny of the Q and M29 haplogroups reconstructed from the consensus network, with Kivisild et al (2005) coding-region sequences added (broken lines). The base changes are transitions unless otherwise indicated, and all polymorphisms shown in regular type are differences with respect to the rCRS. Changes shown in italics are to the same base present in the rCRS. Red type indicates a synonymous change within a protein coding gene, and blue type transitions within the portion of the hypervariable region of the control region used for dating calculations. Changes in RNA genes are indicated by '~r' and '~t' for ribosomal and transfer RNA genes respectively. Details of non-synonymous changes are given in square brackets following the base using three letter protein codes; for example thr-ile describes a change in the codon from threonine to isoleucine. Bases with recurrent (repeat changes at the same base along a single path in the tree) or parallel polymorphisms (changes occurring at a single base more than once in the tree) are marked with an asterisk; and recurrent changes are also indicated by the instance of the 'hit' at that base in the phylogeny shown. The dice symbol indicates that the vertex has been dated (Table 3.2).
The polymorphisms relative to the rCRS at the M vertex are: 73G, 263G, 489C, 750G, 1438G, 2706G, 3106del, 4769G, 7028T, 8701G, 8860G, 9540C, 10398G, 10400T, 10873C, 11719A, 12705T, 14766T, 14783C, 15043A, 15301A, 15326G, 16223T.
Recurrent or parallel polymorphisms in this phylogeny are at nucleotides 59, 92, 152, 208, 1598, 6260, 15172, 16209, 16223, 16241, 16311 and 16519.

variation at the microsatellite region from nt514 to nt523). Two most parsimonious trees were found (from 78 parsimony informative characters, with a parsimony score of 92) by heuristic search in PAUP* (version 4.0b10, Swofford 2003) and this score was proved minimal. The consensus network of the two trees (Figure 3.1a) is itself a tree: the two trees differed only in the order of branching from the Q2 vertex. Figure 3.1b reconstructs the changes along the branches, and includes the coding-only sequences. There was a region of simple conflict in the Q3 subgroup in the Oceanic consensus networks (Chapter Two): this is shown to involve transitions at nt9254 and nt15172, and has been resolved in the labelled tree by invoking a recurrent mutation at nt15172 in the DQ112898 sequence.

The polymorphisms linking the two haplogroups are at nt13500 and nt16311. The nt16311 position appears to be relatively highly mutable (see Chapter Four), and the change to cytosine is not in fact shared by all M/Q and M29 sequences: the Q2 sublineage sequences have a thymine at this position. The nt13500 thymine to cytosine transition also occurs in several other global lineages (in 12 sequences, from haplogroups N/R/F2, M/M7, M/D, N/R/HV, N/R/U and N/R mtDB, Ingman and Gyllensten 2006, search 01/07/07). This analysis supports the suggestion of Merriwether *et al.* (2005) that more M29 sequences are required to confirm the monophyly of Q and M29.

Entire mtDNA samples belonging to the Q haplogroup have been sequenced from individuals from both Near and Remote Oceania; and M29 from samples from New Britain in Near Oceania. In Near Oceania the whole genome sequences obtained have aided in assigning existing control region sequences to haplogroups, providing valuable information about the geographic distributions of the M29 haplogroup and the Q

sublineages (Friedlaender *et al.* 2005, Merriwether *et al.* 2005, Friedlaender *et al.* 2007; see also Chapter Five).

The four M/Q sequences from this study (DQ372884, DQ372885, DQ372880 and DQ372882) are the first reported from Remote Oceania. They form two groups within the Q1 subhaplogroup. The two from Vanuatu (DQ37280 and DQ372882) are closely related, differing at only two positions, as are the two from Polynesia (DQ372884 and DQ372885) which differ at a single site. Both the Vanuatu and Polynesian lineages are linked only by single control region polymorphisms to other Q1 sequences, providing little information about the likely geographic locations of the ancestors of the Remote Oceanic settlers other than Near Oceania. The 13 Q1 sequences have an average of 3 synonymous protein coding changes from the Q1 vertex; using the synonymous transition rate of Kivisild *et al.* (2006) this gives a TMRCA of ~21 300 ± 4400 years (Table 3.2). It is probable that the lineages in Remote Oceania have more recent common ancestry in Near Oceania than is seen from the phylogeny at present; and future targeted studies may be able to identify extant closer relatives of the Remote Oceanic lineages.

Although nt16241G can be seen as a defining HVR-I polymorphism for Q the sequencing of the two Vanuatu sequences DQ372880 and DQ372882 (at the far right of Figure 3.1b) has revealed that this has changed to adenine at some point in the evolution of this sublineage. Q1 sequences have the advantage of a distinctive HVR-I haplotype: they can be readily identified on the basis of nt16144C, nt16148T, nt16265CtvA and nt16343G variants (Figure 3.1b). However, below the Q1 vertex the descendent lineages cannot necessarily be identified solely from the HVR-I sequence, for example the sample from Samoa (DQ372885) has no variants in this portion of the control region from the Q1 vertex. There are however several candidate polymorphisms in the coding region for screening of Q1 sublineages. For example sequencing a portion of the ND1 gene using primers 5F and 5R (Rieder at al 1998, see Appendix B) would cover the nt3394, nt3398 and nt3834 positions and would help in assigning Q1 HVR-I haplotype sequences to a sublineage.

N/R/P, analysed with N/R/R21

The P haplogroup (Figure 3.2) contains sequences from Near Oceania and Australia, and is the only mtDNA haplogroup found in Australia which has common ancestry with any sequences outside of Australia more recently than the L3 'Out of Africa' polytomy. The seven major branches of P share a single synonymous nucleotide transition, at nt15607, in the Cyt. *b* gene, with a probable polytomy directly following this transition. The sublineages have been labelled P1-P7, previously described as independent branches from the P vertex (Friedlaender *et al.* 2005). The descent of the P sequences from the R vertex in the Oceanic consensus networks (Chapter Two) was complicated by conflicting branching possibilities involving the

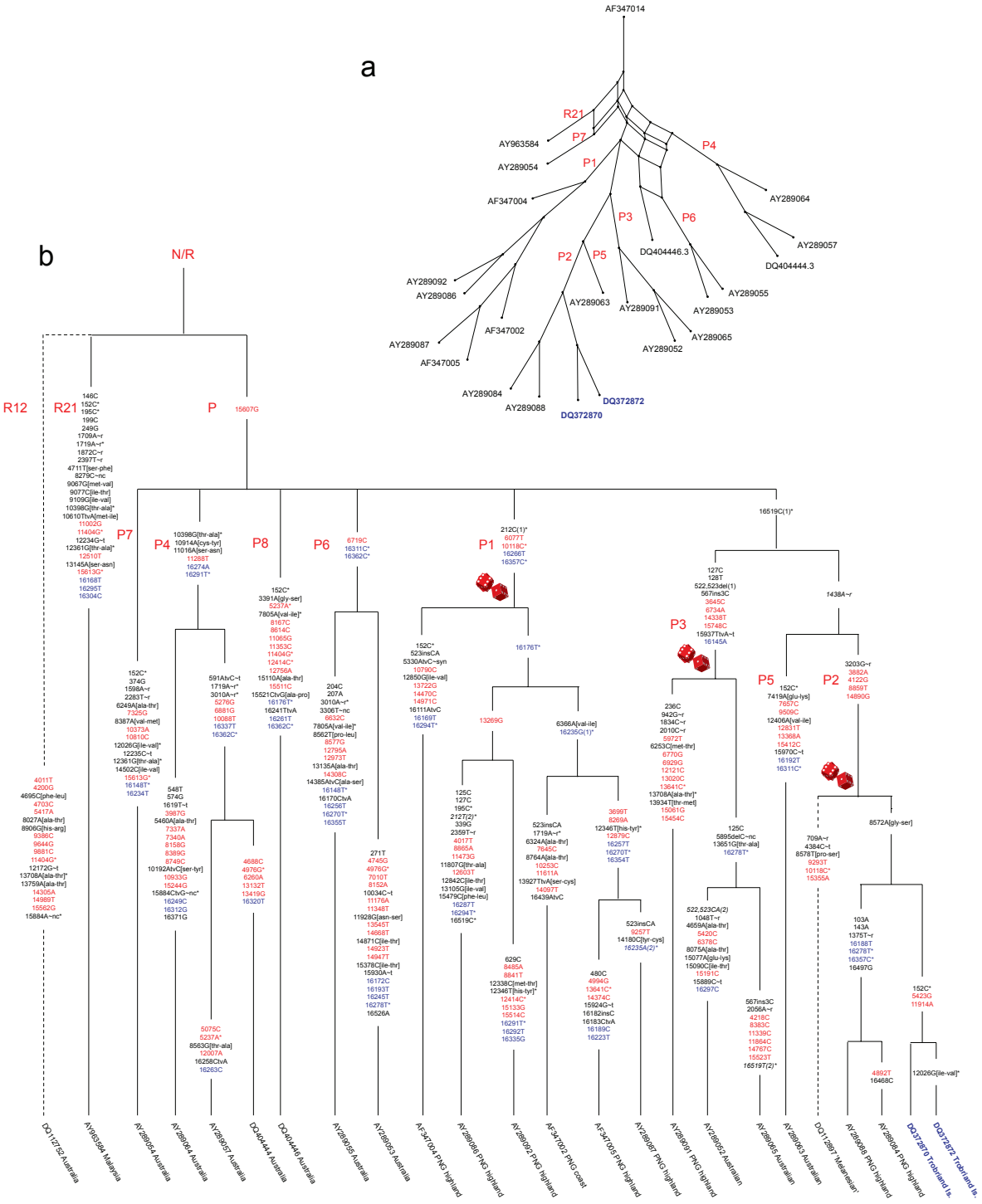


Figure 3.2 P haplogroup with R21 and R12 consensus network and branch-labelled phylogeny
a) The consensus network of 9 most parsimonious trees, from 21 N/R/P individuals with the R21 and AF347014 L1a sequences found by heuristic PAUP* search, (version 4. 0b10, Swofford 2003), constructed using SplitsTree (version 4.6, Huson and Bryant 2006). The entire mtDNA sequence was analysed: there are 74 parsimony informative characters, and the parsimony score of 110 was proved minimal using MMS.
b) A base-labelled phylogeny reconstructed from the consensus tree with Kivisild et al (2006) P and R12 coding-region sequences added (broken lines). See caption for Figure 3.1 for explanation of abbreviations, colours and codes used.
The polymorphisms relative to the rCRS at the N/R vertex are: 73G, 263G, 750G, 1438G, 2706G, 3106del, 4769G, 7028T, 8860G, 11719A, 14766T, 15326G.
Recurrent or parallel polymorphisms in this phylogeny are present at 152, 195, 212, 1719, 4976, 5237, 7805, 10118, 10398 (P and R21), 11404 (P and R12,R21), 12026, 12346, 12361, 12414, 13641, 13708 (P and R12), 15613 (P and R21), 15884 (P and R12), 16148, 16176, 16270, 16278, 16291, 16294, 16311, 16357, 16362 and 16519.

P7 sequence from Australia, and R21 and R12 sequences, from Malaysia and Australia respectively. In the haplogroup analysis the entire R12 sequence was included with the data set, and the coding-only R12 sequence added to the base-labelled phylogeny to explore these links in greater detail.

The consensus network of nine trees found for the P haplogroup sequences, with the single R21 sequence included, displays several conflicting branching hypotheses at the point of common ancestry of the P sublineages and the R21 haplotype (Figure 3.2a). There are two polymorphisms (transitions at nt12361 and nt15613, Figure 3.2b) shared between the single sequences representing lineages R21 and P7, which require for explanation either parallel substitutions at both sites with a MRCA of R21 and P7 at the N/R vertex, or a common ancestor more recently with the P-defining nt15607G transition arising independently in the P7 sequence, or reverting to nt15607A in the R21 sequence. Another shared polymorphism, at nt11404 from A to G, groups the R12 and R21 sequences together. The increased complexity of the branching at this point (seen in Oceanic-133 compared to Oceania-127, Chapter Two, Figures 2.3 and 2.4) is caused by the addition of DQ404446, a recently described Australian sequence (van Holst Pellekaan *et al.* 2006) which also shares the 11404G variant, and appears to branch from the P vertex.

The complexity in the consensus network also involves transitions at nt10398, one of the two homoplasious characters excluded from the Oceanic data set analyses. A transition from guanine to adenine at nt10398 is thought to have occurred along the branch leading to the N vertex from L3; here the three P4 subhaplogroup sequences have an apparent reversion to nt10398G, as does the R21 sequence.

The labelled phylogeny (Figure 3.2b) reveals that the polymorphisms grouping three of the P subhaplogroups ((P2,P5),P3) in the consensus network do not form a particularly strong argument for the structure shown, in contrast to descent of each independently from the P vertex. The nt16519 transition is recognised as highly variable and transitions at nt1438 occur in parallel in several global lineages (see Chapter Four).

In the P haplogroup data set constructed from Oceanic-127, before the addition of the two recent Australian sequences (DQ404446 and DQ404444) only a single tree was found (Supplementary Figure 4, Pierson *et al.* 2006). While DQ404444 falls clearly into the existing P4 subhaplogroup, the placement of DQ404446 is not so straightforward. In the labelled tree (Figure 3.2b) it is drawn as a new sublineage from P as described by van Holst Pellekaan *et al.* (2006) who have named it P8. The link between it and the two P6 sequences seen in the consensus network is at nt16362 (a highly variable HVR-I nucleotide). A second variant, at nt7805, is also shared between DQ404446 and one of the P6 sequences.

The links between Australian and Near Oceanic sequences in the P haplogroup appear deep in the phylogeny. Ten of the 22 P sequences are from Australia, 11 are from Papua New Guinea including 2 from this study from the Trobriand Islands, and a single sequence from Kivisild *et al.* (2006) is described as Melanesian. The sublineages P4, P6, P7, P8 (and also possibly P5), which contain only Australian sequences, branch directly from the P vertex. The Oceanic P sequences belong to the subgroups P1, P2 and P3. P3 is the only sublineage to contain sequences from both Australia and Near Oceania, and the average of ~5.5 synonymous protein changes to the MRCA of the three sequences in this group results in a TMRCA estimate of ~38 400 ± 9300 years (Table 3.2).

As the numbered P subhaplogroups have only a single coding-region polymorphism at nt15607 in common, each has a different HVR-I signature, with many of the polymorphisms occurring in parallel (recurrent or parallel changes occur at these nucleotides: nt16148, nt16176, nt16270, nt16278, nt16291, nt16294, nt16311, nt16357, nt16362). The P haplogroup entire mtDNA sequences present a strong case for the need for additional typing of variants outside of the control region to assign sequences to haplogroups. An example of this is seen in the two sequences from the Trobriand Islands generated by this project (DQ372870 and DQ372872), which are unusual in that their HVR-I haplotype is identical to the rCRS which belongs to the N/R/HV/H haplogroup common to European populations. This raises the possibility that Oceanic P2 samples sequenced only across the HVR-I could be mistakenly assigned to the N/R/HV haplogroup, and assumed to be descended from historic-period immigrants to Oceania.

Recently, Friedlaender *et al.* (2007) have reported seven new entire P sequences, five from New Guinea, one from New Britain and one from Australia, identified as belonging to the sublineages P2, P3 and P4. The two belonging to P3 are from New Guinea and Australia, and reveal a closer link between these regions than previously seen, with four and three synonymous protein changes separating them from their common ancestor. The remaining four New Guinea sequences are described as forming a new sub-group of P4, P4a, with the existing three Australian P4 sequences labelled P4b; representing another link between Australian and Near Oceanic sequences within the P haplogroup. A re-analysis of this haplogroup incorporating the new sequences was not possible due to the time constraints on this study.

M27 and M28 haplogroups

In the Oceanic consensus networks the M27 and M28 haplogroups share a common branch from the M vertex. Together the haplogroups contain 13 samples, and the entire sequences were analysed in a single data set using the MMS approach. A single tree (Figure 3.3a) was found by heuristic search. The two M28 sequences from Vanuatu sequenced from this study (DQ372879 and DQ372883) are the first described from

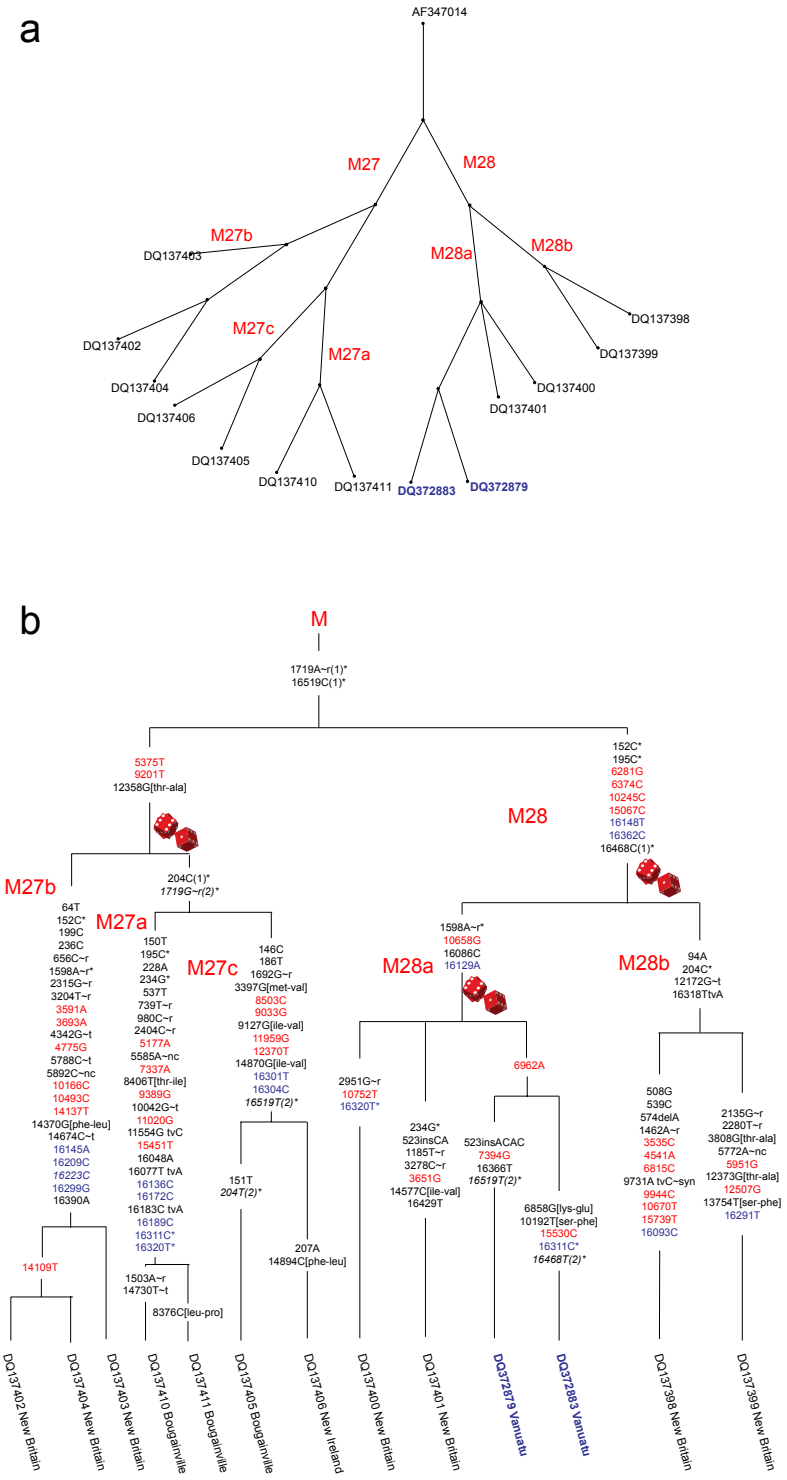


Figure 3.3 M27 and M28 haplogroup minimal tree and labelled phylogeny
a) The single most parsimonious minimal tree (drawn using Spectronet, version 1.25 Huber et al 2002), for 13 M27 and M28 individuals with the AF347014 L3a outgroup found by heuristic PAUP* search (version 4. 0b10, Swofford 2003). The entire mtDNA sequence was analysed. There are 82 parsimony informative characters, and the parsimony score of 95 was proved minimal by MMS. Sequences from this study are shown in blue type.
b) A base-labelled phylogeny reconstructed from the minimal tree showing basal links at nt1719A and nt16519C. See caption for Figure 3.1 for explanation of abbreviations, colours and codes used.
The polymorphisms relative to the rCRS at the M vertex are: 73G, 263G, 489C, 750G, 1438G, 2706G, 3106del, 4769G, 7028T, 8701G, 8860G, 9540C, 10398G, 10400T, 10873C, 11719A, 12705T, 14766T, 14783C, 15043A, 15301A, 15326G, 16223T. Recurrent or parallel polymorphisms in this phylogeny are present at 152, 195, 204, 234, 1598, 1719, 16311, 16320, 16468 and 16519.

Remote Oceania; the remaining 11 samples come from New Britain, New Ireland and Bougainville in Near Oceania. The labelled phylogeny (Figure 3.3b) shows the variants shared between M27 and M28 are nt1719A and nt16519C; bases at which parallel changes occur in several global lineages, and require changes again within M27 (nt1719 and nt16519) and M28a (nt16519).

The connection between the Remote Oceanic from Vanuatu and the M28a sequences from New Britain is quite recent with an average of 1.5 synonymous changes to the MRCA of the four M28a sequences (an estimated time period of ~10 100 ± 4100 years using the Kivisild *et al.* (2006) rate, Table 3.2). Friedlaender *et al.* (2007) have added two new M28a sequences to the existing four, from New Britain and Malaita in the Solomon Islands, which were not included in the data set analysed here due to time constraints. These two new sequences are two synonymous substitutions distant from the M28a vertex, and interestingly one has a transversion at 16265 from adenine to cytosine: this is also seen in the Q1 lineage. As the 16129A and 16148T transitions are also common to both the Q1 and M28 lineages these convergent polymorphisms may be a source of confusion when assigning to haplogroups from HVR-I sequence data.

'Young' Oceanic haplogroups

N/R/B4a haplogroup

The first of the 'young' haplogroups to be described also contains the greatest number of entire sequences from Oceania. Eight of the twenty mt sequences from this project belong within haplogroup N/R/B4a. A total of 46 B4a whole mt sequences are now available, and come from Taiwan, Japan, Korea, Siberia and Oceania (Ingman *et al.* 2000, Ingman and Gyllensten 2003, Mishmar *et al.* 2003, Pierson *et al.* 2006, Starikovskaya *et al.* 2005, Tanaka *et al.* 2004 and Trejaut *et al.* 2005). Figure 3.4a displays the consensus network of the 1274 most parsimonious trees found for this set of sequences, and Figure 3.4b summarises the polymorphisms present in a labelled tree.

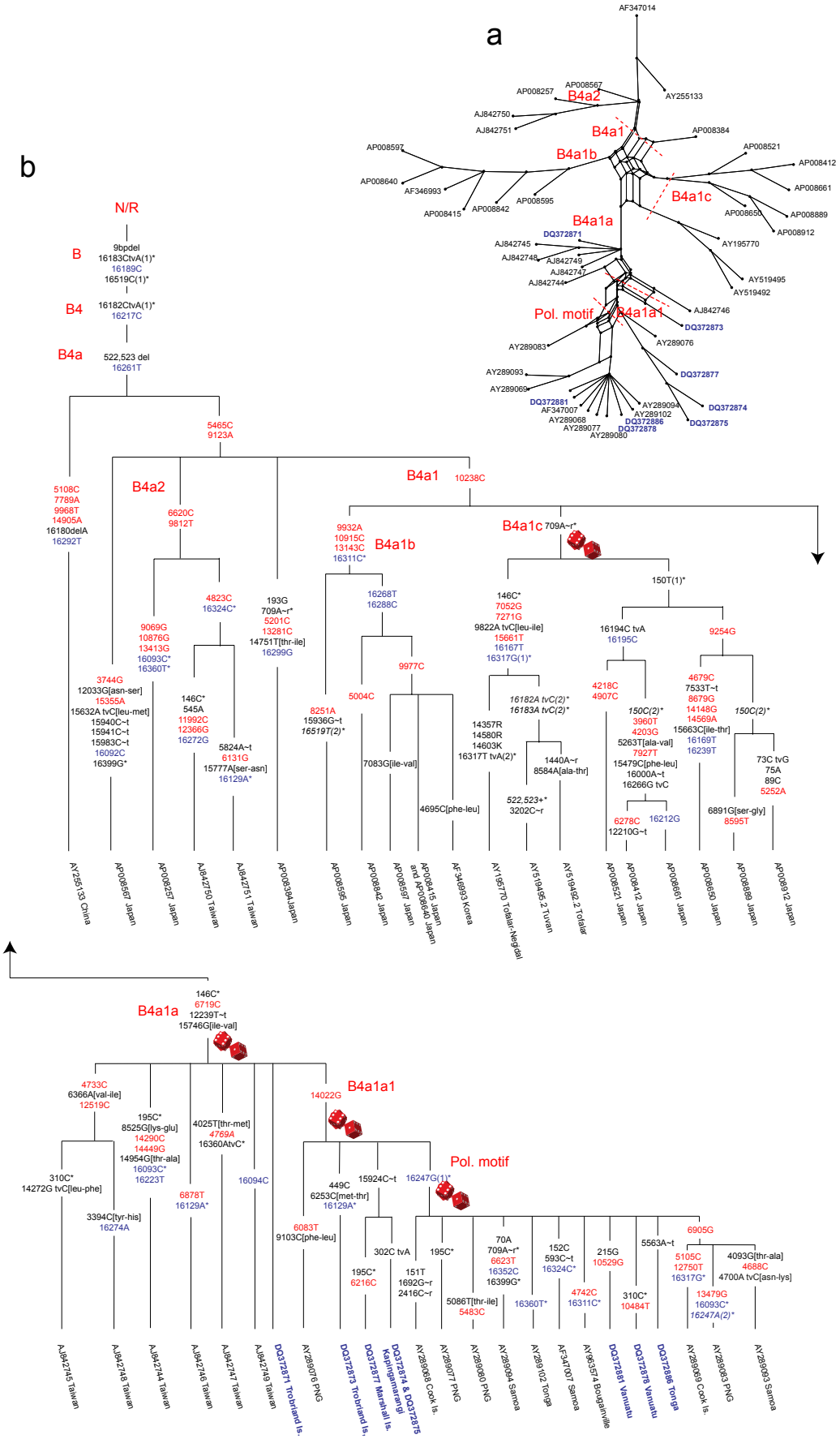
The consensus network displays considerable conflict in branching between the B4a1 subclades and also within the B4a1a sequence set. Substitutions involved in the conflict between the B4a1 subclades are transitions at nt146 and nt709. In the B4a1a sequences three control region polymorphisms and a coding region transition are responsible for the conflict between trees: at nt6905, nt16129, nt16093 and nt16247. In Figure 3.5b the conflict is resolved by effectively weighting coding region polymorphisms over those in the control region, resulting in parallel independent transitions at nt16129 and nt16093 in AJ842746 and DQ372873, and AJ842744 and AY289083. This phylogeny also requires a back mutation event at nt16247 from G to A in AY289083 (this sequence has been checked and confirmed as nt6905G and nt16247A with no

evidence of heteroplasmy; email communication Max Ingman 26/07/06).

All of the Oceanic sequences within haplogroup B4 belong to the B4a1a subgroup, which share three coding region transitions (at nt6719, nt12239 and nt15746) distinguishing them from other B4a1 sequences. A Trobriand Island sequence, DQ372871, has no subsequent changes from these at the B4a1a vertex. The only sequences in this group apart from those from Oceania are six from Taiwan, which are separated from all the Oceanic sequences, with the exception of DQ372871 by the nt14022G transition (Figure 3.4b), which defines the B4a1a1 subgroup. Three sequences from Kapingamarangi Atoll and the Marshall Islands in Micronesia share a transition at nt15924G, perhaps providing a marker for a Micronesian subgroup of B4a1a1. Thirteen (including AY289083; back mutation assumed) of the eighteen sequences in B4a1a1 have the 'Polynesian motif' nt16247G transition, and these sequences come from both Near and Remote Oceania. The only shared variant subsequent to the nt16247G transition is at nt6905; found in sequences from the Cook Islands, Papua New Guinea and Samoa.

In a comprehensive recent summary of the East Asian mtDNA haplogroups, Kong *et al.* (2006) present a revised skeleton phylogeny for all M and N haplogroups present, outlining defining polymorphisms and names given to vertices. These trees are reproduced here (by permission of Oxford University Press) as supplementary Appendix Figures D3.5 and D3.6. In Figure 3.4 I have left some branches unnamed for overall clarity (for example, the far left branch descending from B4a1a has been named B4a1a2 by Kong *et al.* (2006)): the labelled trees contain a large amount of information and I have aimed to present as much detail as possible while still maintaining reasonable print legibility. The tree in Figure 3.4b differs from that presented by Kong *et al.* (2006) in the naming of the vertices B4a1a1 and what I have termed the 'Pol. motif' vertex. Kong *et al.* define B4a1a1 by both the transitions at nt14022 and nt16247, and name the sequences with the nt6905 transition (the far right branch from the 'Pol. motif' vertex in Figure 3.4b) B4a1a1a. This terminology has been adopted by Hill *et al.* (2007), but as I believe more sequences are required to clarify the

Figure 3.4 B4a consensus network and branch-labelled phylogeny (following page)
a) The consensus network of 1274 most parsimonious trees (57 parsimony informative characters over complete mtDNA sequence, parsimony score 87) found by heuristic search (PAUP* version 4.0b10, Swofford 2003) on the N/R/B4a haplogroup data set of 47 sequences including L3c outgroup (AF347014). The parsimony score was proved minimal by MMS and the network was constructed using Spectronet (version 1.25 Huber et al 2002). Sequences in B4a2 are from Taiwan and Japan; B4a1b sequences are from Japan and Korea and B4a1c sequences are from Japan and Siberia. All B4a1a sequences are from Taiwan and Oceania. Sequences from this study are shown in blue type.
b) A base-labelled phylogeny reconstructed from the consensus network for B4a sequences. See caption for Figure 3.1 for explanation of abbreviations, colours and codes used. The polymorphisms relative to the rCRS at the N/R vertex are: 73G, 263G, 750G, 1438G, 2706G, 3106del, 4769G, 7028T, 8860G, 11719A, 14766T, 15326G. The conflicts between the trees seen in the B4a1a subset of the sequences have been resolved here by invoking a reversion at nt16247 from G to A in sequence AY289093. Other positions causing conflict in the phylogeny shown are nt16129 (AJ842746 and DQ372873) and nt16093 (AJ842744 and AY289083). Recurrent or parallel polymorphisms in this phylogeny are present at 146, 150, 195, 310, 709, 16093, 16129, 16182, 16183, 16247, 16311, 16317, 16324, 16360 and 16399.


a
AF347014
AP008567
AP008257
B4a2
AY255133
AJ842750
AJ842751
B4a1
AP008384
AP008597
B4a1b
AP008521
AP008412
AP008640
AF346993
AP008595
B4a1c
AP008661
AP008415
AP008842
AP008650
AP008889
B4a1a
DQ372871
AP008912
AJ842745
AY195770
AJ842748
AJ842749
AJ842747
AY519495
AJ842744
AY519492
Pol. motif
B4a1a1
AJ842746
DQ372873
AY289083
AY289076
DQ372877
AY289093
AY289069
DQ372881
AF347007
AY289094
DQ372874
AY289102
AY289068
DQ372886
DQ372875
AY289077
DQ372878
AY289080
b
N/R
B
9bpdel
16183CtvA(1)*
16189C
16519C(1)*
B4
16182CtvA(1)*
16217C
B4a
522,523 del
16261T
5465C
9123A
5108C
7789A
9968T
14905A
16180delA
16292T
B4a2
6620C
9812T
B4a1
10238C
9932A
10915C
13143C
16311C*
B4a1b
B4a1c
709A~r*
16268T
16288C
150T(1)*
4823C
16324C*
193G
709A~r*
5201C
13281C
14751T[thr-ile]
16299G
9069G
10876G
13413G
16093C*
16360T*
146C*
7052G
7271G
9822A tvC[leu-ile]
15661T
16167T
16317G(1)*
16194C tvA
16195C
9254G
3744G
12033G[asn-ser]
15355A
15632A tvC[leu-met]
15940C~t
15941C~t
15983C~t
16092C
16399G*
146C*
545A
11992C
12366G
16272G
5824A~t
6131G
15777A[ser-asn]
16129A*
9977C
5004C
8251A
15936G~t
16519T(2)*
7083G[ile-val]
4695C[phe-leu]
16182A tvC(2)*
16183A tvC(2)*
14357R
14580R
14603K
16317T tvA(2)*
1440A~r
8584A[ala-thr]
522,523+*
3202C~r
4218C
4907C
150C(2)*
3960T
4203G
5263T[ala-val]
7927T
15479C[phe-leu]
16000A~t
16266G tvC
6278C
122106~t
16212G
4679C
7533T~t
8679G
14148G
14569A
15663C[ile-thr]
16169T
16239T
150C(2)*
73C tvG
75A
89C
5252A
6891G[ser-gly]
8595T
AY258133 China
AP008567 Japan
AP008257 Japan
AJ842750 Taiwan
AJ842751 Taiwan
AP008384 Japan
AP008595 Japan
AP008842 Japan
AP008597 Japan
AP008415 Japan and AP008640 Japan
AF346993 Korea
AY195770 Tobilar-Negrital
AY519496.2 Tuvan
AY519492.2 Tofalar
AP008521 Japan
AP008412 Japan
AP008661 Japan
AP008650 Japan
AP008889 Japan
AP008912 Japan
B4a1a
146C*
6719C
12239T~t
15746G[ile-val]
4733C
6366A[val-ile]
12519C
195C*
8525G[lys-glu]
14290C
14449G
14954G[thr-ala]
16093C*
16223T
4025T[thr-met]
4769A
16360AtvC*
B4a1a1
14022G
310C*
14272G tvC[leu-phe]
3394C[tyr-his]
16274A
6878T
16129A*
16094C
449C
6253C[met-thr]
16129A*
15924C~t
16247G(1)*
Pol. motif
6083T
9103C[phe-leu]
195C*
6216C
302C tvA
151T
1692G~r
2416C~r
195C*
5086T[thr-ile]
5483C
70A
709A~r*
6623T
16352C
16399G*
16360T*
152C
593C~t
16324C*
4742C
16311C*
215G
10529G
310C*
10484T
5563A~t
6905G
5105C
12750T
16317G*
13479G
16093C*
16247A(2)*
4093G[thr-ala]
4688C
4700A tvC[asn-lys]
AJ842745 Taiwan
AJ842748 Taiwan
AJ842744 Taiwan
AJ842746 Taiwan
AJ842747 Taiwan
AJ842749 Taiwan
DQ372871 Trobriand Is.
AY289076 PNG
DQ372873 Trobriand Is.
DQ372877 Marshall Is.
DQ372874 & DQ372875 Kapingamarangi
AY289068 Cook Is.
AY289077 PNG
AY289080 PNG
AY289094 Samoa
AY289102 Tonga
AF347007 Samoa
AY963574 Bougainville
DQ372881 Vanuatu
DQ372878 Vanuatu
DQ372886 Tonga
AY289069 Cook Is.
AY289083 PNG
AY289093 Samoa

phylogeny below the nt14022G variant it is not followed here.

The entire mtDNA sequences reveal considerable distance between the occurrence of the HVR-I 'Polynesian motif' (PM) nt16261T and nt16247G transitions. The nt16261T transition is common to all of the B4a subgroup, which includes sequences from Japan, China, Siberia, Korea and Taiwan as well as Oceania: this point in the phylogeny is separated from the N/R vertex only by HVR-I polymorphisms and the occurrence of the 9 base-pair deletion (Figure 3.5b). Seven coding region transitions (five synonymous transitions) separate this point and the occurrence of the nt16247G in the Oceanic subgroup. This is important when interpreting available HVR-I sequences: the pre-PM and PM sequences are not as closely related as has been assumed.

The TMRCA estimates in this haplogroup are of particular interest (Table 3.2) as its current distribution is so closely associated with theories of recent migration into Oceania. The B4a1a subgroup includes sequences from Taiwan and Oceania, and the date estimate of the most recent common ancestor of this group of 25 sequences taken from synonymous coding region changes is ~10800 ± 1800 years. The B4a1a1 and 'Polynesian Motif' subgroups nested within B4a1a have TMRCA estimates of ~5300 ± 1400 and ~6200 ± 1800 years respectively. These estimates are not considerably greater than archaeological dates for the first Lapita settlements in Near Oceania, in contrast to HVR-I TMRCA estimates (Richards *et al.* 1998) that suggested an Island Southeast Asian origin of the Polynesian Motif sequence at ~ 18000BP.

The N/R/B4a entire mtDNA phylogeny provides cues for future research: typing individuals from Island Southeast Asia and the Philippines with the HVR-I pre-PM for their residues at the coding region nucleotides defining B4a1a1 (14022) and B4a1a (6719, 12239, 15746) would make a considerable contribution to our understanding of the development and spread of this haplogroup into Oceania. Within Oceania, it may be possible to trace particular sub-types, for example, using the nt15924 transition found in the Micronesian samples described here, to elucidate prehistoric patterns of migrations and post-settlement interactions.

M/M7bc, analysed with M22

Two samples sequenced for this study belong to the M7bc haplogroup, which has entire mtDNA sequences reported from Japan, Taiwan, Mongolia, China, and the Philippines, and now from Micronesia in Remote Oceania. Figure 3.5a shows the consensus network constructed from the 5040 most parsimonious trees found: most of the conflict between the trees occurs in the subset of sequences circled in the M7b part of network. These are closely related haplotypes, with all but one derived from Japan, and the conflict appears to be based around control region substitutions at nt16183 adjacent to the polyC stretch (personal observation of the 16189T to C transition which occurs along the branch leading to these sequences in several global

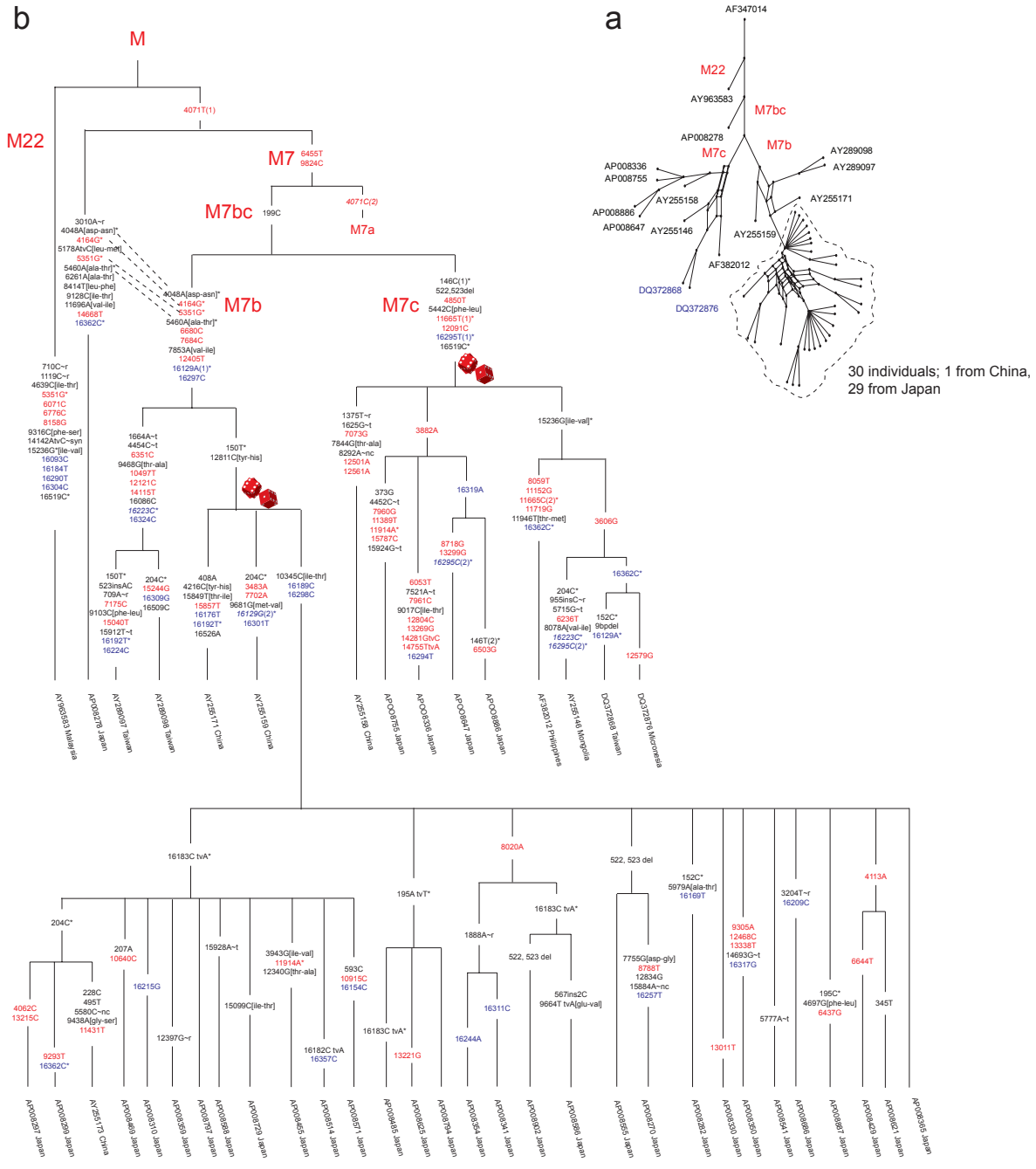


Figure 3.5 M/M7bc with M22 consensus network and labelled phylogeny

a) The consensus network of 5040 most parsimonious trees (constructed using Spectronet version 1.25 Huber et al 2002) found by heuristic PAUP* search (version 4.0b10, Swofford 2003), from 44 M7bc individuals with the M22 and AF347014 L3a sequences. The entire mt genome sequence was analysed; there are 55 parsimony informative characters and the parsimony search score of 89 was proved optimal by MMS. Sequences from this study are shown in blue type. b) A labelled reconstruction of the M/M7bc phylogeny. Dotted lines between AP008278 and M7b sequences highlight shared polymorphisms. See caption for Figure 3.1 for explanation of abbreviations, colours and codes used. The polymorphisms relative to the rCRS at M vertex are 73G, 263G, 489C, 750G, 1438G, 2706G, 3106del, 4769G, 7028T, 8701G, 8860G, 9540C, 10398G, 10400T, 10873C, 11719A, 12705T, 14766T, 14783C, 15043A, 15301A, 15326G and 16223T. Recurrent or parallel polymorphisms in this phylogeny are at nucleotides 146, 150, 152, 195, 204, 4048, 4164, 5351, 5460, 11665, 11914, 15236 (M22 and M7c), 16129, 16183, 16192, 16223, 16295, 16362 and 16519 (M22 and M7c).

lineages shows it is often followed by transversions at the 16182 and 16183 adenine bases).

The AP008278 sequence, from Japan, which falls outside of the named M7bc haplogroup in Figure 3.5 is unusual, as it has four transitions in common with the M7b sequences (drawn with dotted lines in Figure 3.5b). The nt4071T transition shown in the phylogeny above the M7bc vertex is elsewhere placed with the 199C transition at the M7bc vertex; M7a sequences do not have this variant (Kong *et al.* 2006 Appendix D3.6). This is represented on the tree in Figure 3.5b as a recurrent mutation leading to the M7a subhaplogroup (as these sequences were not included in the analysis and this retains the branching of the consensus network in Figure 3.5a); however it seems probable that the AP008278 sequence contains errors: Kong *et al.* (2006:2083) note that 13 of the 672 sequences from the Tanaka *et al.* (2004) dataset, from which AP008278 derives show evidence of artificial recombination.

In the Oceanic consensus networks shown in Chapter Two, the M7 sequences branch from the M vertex with the single M22 sequence from Malaysia (Macaulay *et al.* 2005). When the additional M7bc sequences from Japan, Mongolia and China are included, and the entire mtDNA sequence analysed, this connection is not seen (Figure 3.5). The coding region variants which link M22 to the M7 sequences in the Oceanic data set are shown in Figure 3.5b; a transition at nt5351 is shared between the M22 sequence and the two Taiwanese M7b sequences, and a transition at nt14236 is common to the M22 sequence and one of the three descendent lineages from the M7c vertex.

The samples from Taiwan and the Marshall Islands sequenced for this study reveal a close relationship between the two M7c haplotypes (Figure 3.5b), suggesting a fit to a model of migration from Taiwan of peoples carrying M/M7c and ancestral N/R/B4a1a haplotypes. The other Taiwanese sequences belonging to M7bc show a much more distant relationship to the Micronesian sample, with common ancestry at the M7bc vertex. However, the sample from the Philippines, which would be expected to show close ties to the Taiwanese and Remote Oceanic sample, has several independent polymorphisms, and is definitely not intermediate between the Taiwanese and Micronesian haplotypes.

The identification of M7c haplotypes from HVR-I sequences is problematic, as from the whole mtDNA phylogeny the differences to the rCRS are expected to be at nt16223 and nt16295; both of which are very common variants and additionally at nt16362, also highly variable. Screening samples for the nt3606 and 15236 transitions would enable members of this haplogroup to be identified for further sequencing to resolve the Oceanic part of this phylogeny and study its distribution in Oceanic populations.

N/R/B5a haplogroup

While it does not contain sequences from Oceania, this haplogroup is included here as a sample sequenced for this study from Taiwan belongs to B5a, a little-known haplogroup for which only ten entire sequences are presently available (Kong *et al.* 2003, Pierson *et al.* 2006, Tanaka *et al.* 2006, Thangaraj *et al.* 2005). In the Oceanic consensus networks (Chapter Two) the Taiwanese sample branched from the B vertex with sequences from the Nicobar Islands. When B5a samples from China and Japan were included, and the whole mtDNA sequence analysed a single most parsimonious tree was found (Figure 3.6a). The labelled tree (Figure 3.6b) details the polymorphisms present in the sequences. The sequences branch from the B5a vertex into two groups; one (B5a1, Appendix D3.5, Kong *et al.* (2006)) containing a sample from China and the four closely-related sequences from the Nicobar Islands and the other (B5a2) the sample from Taiwan, and four Japanese

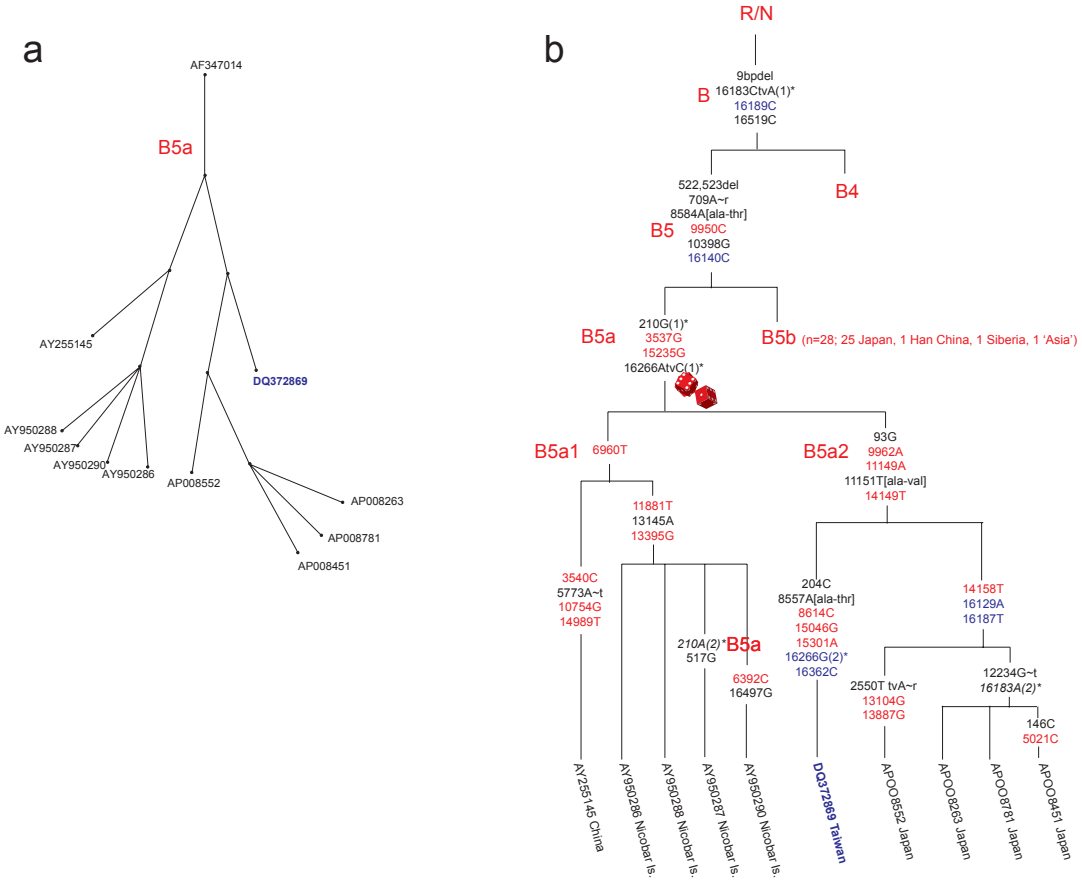


Figure 3.6 B5a haplogroup minimal tree and branch-labelled phylogeny
a) The single most parsimonious tree, found by heuristic PAUP* search (version 4.0b10, Swofford 2003) from 10 B5a individuals with L3 AF347014 outgroup. The entire mtDNA sequence was included; there are 17 parsimony informative characters and the parsimony score of 21 was proved minimal using MMS. The tree was drawn using Spectronet (version 1.25 Huber et al 2002) and the sequence from this study is shown in blue type.
b) A labelled reconstruction of the N/R/B5a phylogeny. See caption for Figure 3.1 for explanation of abbreviations, colours and codes used. The polymorphisms relative to the rCRS at the N/R vertex are: 73G, 263G, 750G, 1438G, 2706G, 3106del, 4769G, 7028T, 8860G, 11719A, 14766T, 15326G. Recurrent or parallel polymorphisms in this phylogeny are at nucleotides 210, 16183 and 16266.

sequences.

3.3 Issues in dating human mtDNA phylogenies

Table 3.2 lists the dating estimates obtained from the haplogroup analyses. The haplogroups selected from populations outside of the Pacific were chosen for various reasons: N/W (Appendix D3.1) was included initially in this study as a W sequence was obtained from the project, and the W haplogroup, with 27 entire and 10 coding-only sequences is relatively small by comparison to others common in European populations (see Chapter Four). The N/S haplogroup (Appendix D3.2) found only in Australia was selected as it also has a manageable number of sequences but contains the largest number of sequences of the Australian haplogroups (10 entire and two coding-only sequences). The high number of homoplasious mutations amongst the N/R/P sequences, many of which are from Australia, also inspired the choice of this haplogroup as I wanted to examine the two haplogroups for any evidence of shared polymorphisms.

The N/R/B4bd/b subset (Appendix D3.3) was compiled to test the results of dating methods at estimating the most recent common ancestor of the available American haplogroup B sequences and their closest relatives in Asia, and is the only data set in this chapter to exclude the control region. This was necessary as the majority of the sequences were coding-region only (17 of a total 23). As all of the other data sets branch from the 'Out of Africa' M and N (and N/R) vertices of the human mtDNA tree, an older haplogroup, L1c, (Appendix D3.4) was analysed to compare the dating methods over a longer time-frame.

While in several instances the number of sequences contributing to the TMRCA estimate is very small (for example only three sequences for M/Q2 and N/R/P3), 10 of the 20 vertices dated using the three sequence subsets and rho methods have 10 or more descendent sequences.

The dates obtained from the different calculations for each vertex vary considerably, although in most cases the estimates obtained from both coding methods overlap when the standard errors are taken into account. However the M27, M7b2, L1c and N/S vertices do not show this pattern. By contrast, comparing the synonymous transition estimates to those from the HVR-I calculation for each vertex shows little overlap, with the estimates from the control region often greater than the synonymous rate estimations by a factor of 2 or more. To further examine the relationship between the rate of non-synonymous transitions, coding region changes and HVR-I transitions in general, Table 3.3 lists the ratios of the average numbers of synonymous transitions to coding region changes and HVR-I transitions to synonymous transitions, for vertices with 10 or more descendent sequences.

Table 3.2 TMRCA estimates

		Synonymous Transitions[a]			Coding Region nt577–nt16022[b]			HVR-I Transitions nt16090-nt16365[c]			
Vertex	***n***	***ρ***	***σ***	***ρ ± σ* (years)**	***ρ***	***σ***	***ρ ± σ* (years)**	***n***	***ρ***	***σ***	***ρ ± σ* (years)**
Q	22	4.95	0.56	33 482 ± 3788	8.59	0.72	44 135 ± 3699	16	3.38	0.46	68 208 ± 12 108
Q1&Q2	16	5.13	0.70	34 699 ± 4735	7. 36	0.83	39 203 ± 4265	13	3.77	0.54	76 078 ± 10 897
Q1	13	3.15	0.65	21 307 ± 4387	5.15	0.80	26 460 ± 4110	10	1.70	0.41	34 306 ± 8274
Q2	3	3.00	1.00	20 292 ±6764	4.67	1.25	23 994 ± 6423	3	0.33	0.33	6659 ± 6659
P1	6	4.83	0.90	32 670 ± 6088	8.17	1.17	41 977 ± 6011	6	4.00	0.82	80 720 ± 16 547
P2	5	1.60	0.57	10 822 ± 3855	3.60	0.85	18 497 ± 4367	4	1.50	0.61	30 270 ± 12 310
P3	3	5.67	1.37	38 352 ± 9267	10.67	1.89	54 822 ± 8711	3	1.00	0.58	20 180 ± 11 704
M27	7	4.71	0.94	31 858 ± 6358	13.86	1.62	71 418 ± 8221	7	3.71	0.83	74 868 ± 16 749
M28	6	3.00	0.71	20 292 ± 4802	6.33	1.03	32 524 ±5292	6	1.50	0.50	30 270 ± 10 090
M28a	4	1.50	0.61	10 146 ± 4126	3.00	0.87	15 414 ± 4470	4	0.50	0.35	10 090 ± 7063
B4a1c	9	3.00	0.58	20 292 ± 3923	4.56	0.71	23 429 ± 4735	9	1.33	0.38	26 839 ± 7668
B4a1a	25	1.60	0.26	10 822± 1759	2.42	0.31	12 331 ± 1593	25	1.04	0.20	20 987 ± 4036
B4a1a1	18	0.78	0.21	5276 ± 1429	1.50	0.30	7707 ± 1541	18	1.17	0.25	23 611 ± 5045
'Pol. Motif'	13	0.92	0.27	6223 ± 1826	1.54	0.34	7913 ± 1747	13	0.54	0.20	10 897 ± 4036
M7c	9	3.00	0.58	20 292 ± 3923	5.11	0.75	26 255 ± 3854	9	1.11	0.35	22 400 ± 7063
M7b2	32	0.75	0.15	5073 ± 1015	2.34	0.27	12 023 ± 1541	32	2.31	0.27	46 616 ± 5449
N/B5a	10	4.10	0.74	27 732 ± 5005	5.70	0.88	29 287 ± 4521	10	1.00	0.37	20 180 ± 7467
N/W	36	2.06	0.24	13 934 ± 1623	4.03	0.33	20 706 ± 1541	26	0.19	0.09	3834 ± 1816
N/S	12	3.00	0.50	20 292 ± 3382	7.75	0.80	39 820 ± 4110	9	2.11	0.48	42 580 ± 9686
N/R/B4b	23	5.09	0.47	34 429 ± 3179	8.57	0.61	44 033 ± 3134				
B4b 'Americas'	19	2.47	0.36	16 707 ± 2435	4.11	0.46	21 117 ± 2363				
L1c	29	12.44	0.88	84 144 ± 5952	20.88	1.14	107 281 ± 5857	16	3.00	0.43	60 540 ± 8677

[a] One synonymous transition per 6764 years (Kivisild et al. 2006).
[b] One substitution per 5138 years (Mishmar et al. 2003).
[c] One transition per 20 180 years (Forster et al. 1996).

As indicated by the date estimates, the average number of synonymous transitions per coding region change in general is not strikingly different at each of the 10 vertices; with all but the M7b2 vertex showing an average of between 0.51 and 0.72 synonymous transitions per substitution seen across the entire coding region, and little differentiation between the older haplogroups such as L1c and Q, and the more recent vertices. The synonymous transition rate tends to produce more conservative time estimates, and has the advantage over calculations based on all coding region substitutions in that the variation studied has no immediately apparent biological consequence, and may not be subject to the same selective forces as changes which result in amino acid substitutions or occur in RNA genes.

The ratio of the average number of HVR-I transitions to the average synonymous transitions ranges from 0.09 changes in the HVR-I region for each synonymous transition for vertex N/W, to 3.08 HVR-I changes per synonymous transition in the case of vertex M7b2. The haplogroup with the greatest distance to common ancestry, L1c, shows a ratio of 0.24 HVR-I transitions to synonymous transitions. While five of the ratios fall between 0.59 and 0.73 there does not appear to be a clear relationship between synonymous and HVR-I

transitions, suggesting that the rate of change of nucleotides in this region is not constant with respect to the changes in the coding region.

Re-calibration of the HVR-I transition rate by comparison to the synonymous transition rate estimate with such a wide range of variation does not seem practical. The five ratios of HVR-1 to synonymous changes which vary between 0.59 and 0.7 have a mean of 0.64, equating to an HVR-I rate of 1 transition per 10 569 years (given the synonymous transition rate of 1 per 6764 years). However the large range of average changes seen in the phylogenies suggests that the rates of change of several nucleotides in the HVR-I, often termed hypervariable, need to be taken into account in the calculation of average changes to an ancestral vertex. In Chapter Four the rates of change in the control region are examined more explicitly.

The HVR-I transition rate of 1 transition per 20 180 years between nucleotides 16090 and 16365 calculated by Forster *et al.* (1996) was based on variation in sequences belonging to haplogroup A in North American Eskimo and Na-Dene populations, and calibrated using a date of 11 300 years for the Younger Dryas glacial relapse for the most recent common ancestry of the contemporary haplotypes. A later study by the same group (Saillard *et al.* 2000, in a section entitled 'Reassessment of the mtDNA mutation rate') discusses problems with this calculation due to the suboptimal quality of some of the sequence data included, concluding that some polymorphisms were probable artefacts. While the problems with the rate

Table 3.3 Ratios of rho for vertices with more than 10 descendant sequences

		Synon	Coding	Synon/ Coding		HVR-I	HVR-I/ Synon
Vertex	**n**	***ρ***	***ρ***		**n**	***ρ***	
Q	22	4.95	8.59	0.58	16	3.38	0.68
Q1&Q2	16	5.13	7.36	0.70	13	3.77	0.73
Q1	13	3.15	5.15	0.61	10	1.70	0.54
B4a1a	25	1.60	2.42	0.66	25	1.04	0.65
B4a1a1	18	0.78	1.50	0.52	18	1.17	1.50
'Pol. Motif'	13	0.92	1.54	0.60	13	0.54	0.59
M7b2	32	0.75	2.34	0.32	32	2.31	3.08
N/R/B5a	10	4.10	5.70	0.72	10	1.00	0.24
N/W	36	2.06	4.03	0.51	26	0.19	0.09
L1c	16	12.44	20.88	0.60	16	3.00	0.24

Table 3.4 Coding-region substitution rates from Atkinson (2006)

	Calibration (kyrs)			Est. sub. rate (1 per x years)		
	lower	**mean**	**upper**	**lower**	**mean**	**upper**
N/R/B4a1	3	4	5	1660	912	629
M/Q	40	50	60	6474	4316	3237
Human-Chimp	6000	6500	7000	5395	4980	4316

are acknowledged, the authors continue to use it in the later study 'to facilitate comparison with previously published age estimates, analogous to uncalibrated radiocarbon years' (Saillard *et al.* 2000:722). In several studies where the rho dating method for HVR-I sequences has been used without discussion of issues involving the rate calculation, both Forster *et al.* (1996) and Saillard *et al.* (2000) are cited (for example Abu-Amero *et al.* 2007, Hill *et al.* 2007, Melton *et al.* 2007, Roostalu *et al.* 2007, Rowald *et al.* 2007: there are 92 citations for Saillard *et al.* 2000 and only papers from 2007 are listed here). With the accumulation of large numbers of entire human mtDNA sequences in recent years, and development of new dating techniques (Atkinson 2006, Drummond *et al.* 2006), a re-evaluation of the utility of this dating method would be timely.

The calibration point used to calculate both the synonymous transition and all coding substitution rates is the most recent common ancestor of chimp and human, estimated at 6.5 million years ago (Goodman *et al.* 1998). These rates were presumably calculated from the average human mtDNA-chimpanzee distances using a variety of mutational models (Mishmar *et al.* 2003 state the HKY85 model was used; in Kivisild *et al.* 2006 details are not given). Computational limitations have in the past restricted the use of likelihood analyses with complex models of DNA evolution to small data sets. Recently however, Atkinson (2006) has used a Bayesian approach to investigate the age of mitochondrial Eve under a likelihood framework which incorporates multiple calibration points.

The rates of change in the coding region of a global data set of 252 human, two chimpanzee and one gorilla sequences were estimated using two models, the first a General Time Reversal (GTR) substitution model allowing gamma distributed rates and a proportion of invariant sites and the second incorporating codon-specific rates in a GTR substitution model. Control region sequences were analysed separately using the GTR model allowing gamma distributed rates and a proportion of invariant sites. The calibration points

used were lower, mean and upper estimates for the TMRCA of human and chimpanzee sequences, the 11 M/Q sequences from Friedlaender *et al.* (2005), Ingman *et al.* (2000) and Ingman and Gyllensten (2003) and the 11 B4a1a sequences from Ingman (2000), Ingman and Gyllensten (2003) and Macaulay *et al.* (2005). Table 3.4 reproduces the rate results from this study (Atkinson 2006; adapted from Table 7.3, p7.12) for the first coding region model, allowing comparison to Mishmar *et al.*'s (2003) rate estimate of 1 coding region substitution per 5138 years.

These results illustrate the importance of the calibration point chosen in rate determination, and the increase in the rate of change obtained from the most recent calibration point (for the N/R/B4a1a sequences) raises the question of time dependency of molecular rates (Ho *et al.* 2005, Penny 2005). In conclusion, it seems dates obtained from human phylogenies need to be interpreted with caution, particularly those obtained from HVR-I sequence data from groups with recent common ancestry.

The entire mt genome phylogenies for the Oceanic haplogoups presented here provide a resource for future analyses of sample sets which will be able to target diagnostic coding region polymorphisms to assign sequences reliably to haplogroups. Sourcing a likely Near Oceanic geographic origin of the ancient haplogroup lineages present in Remote Oceania is an exciting potential application of this technique. In Chapter Six, the SNPs identified by entire mt genome sequence of N/R/B4a haplotypes (Figure 3.4) are used to supplement the data gained from HVR-I sequencing of a new set of Polynesian samples, greatly improving the resolution of the data.

The phylogenies described here revealed surprisingly high instances of recurrent mutations (for example N/R/P, Figure 3.2), and in the following chapters the nature of mitochondrial DNA, and its variation in humans, is examined, and analyses carried out to assess the extent and effects of recurrent mutations in phylogenetic reconstructions.

Chapter 4. Variation in Human mtDNA

The phylogenetic analyses of the Oceanic data set and the haplogroup subsets described in Chapters Two and Three revealed many instances of homoplasy - changes at the same base in parallel on different paths from a common ancestor to descendants, or repetitive changes at the same base along a single path. Mutation rate variability within the control region has long been recognised, with certain nucleotides identified as 'hypervariable' (Meyer *et al.* 1999, Stoneking 2000). Understanding the mechanisms underlying the fixation of nucleotide changes in the mtDNA genome is of clear importance to the interpretation of phylogenies generated from sequences. Some positions within the mitochondrial genome may be more vulnerable than others to mutation, and selective effects and recombination events may also contribute to the mutation rate variability observed. The large numbers of entire human mitochondrial genomes now available on public databases are a valuable empirical resource for the exploration of factors contributing to the accumulation of mitochondrial genetic variability.

This chapter begins with a brief review of the current understanding of mtDNA structure, function, replication, and inheritance, and a discussion of prior analyses of selection and recombination in human mtDNA. This is followed by a description of the variation observed in the large data set of global mt sequences assembled over the course of this project, and the results of selection and recombination tests performed on these data. In Chapter Five subsets of the global data set are used in a phylogenetic analysis which explores the occurrence of homoplasy in human mtDNA in greater detail, focusing on the control region in particular. There are several hundred control region sequences available from Oceanic populations, and in Chapter Six the information contained within these sequences is reassessed, making use of the resolution gained from the whole mt genome analyses (Chapters Two and Three) and the insights into rate variability obtained from Chapter Five.

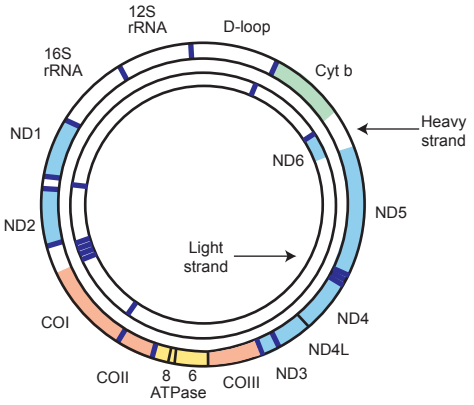
4.1 Introduction

Structure and function of mtDNA

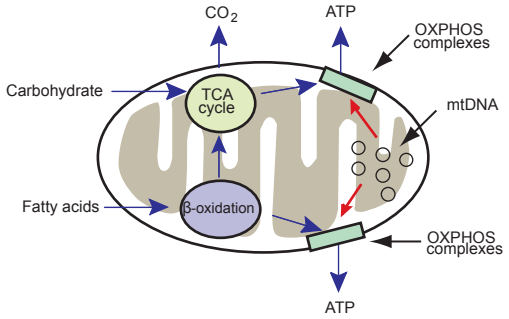
The mtDNA molecule encodes two ribosomal RNA (rRNA) molecules, 13 proteins and 22 transfer RNA (tRNA) molecules. The tRNAs and rRNAs are required to synthesise the proteins which are then incorporated into four of the five multiprotein enzyme complexes of the mitochondrial oxidative phosphorylation (OXPHOS) system. The OXPHOS system, which also contains many known subunits encoded by nuclear genes, generates adenosine triphosphate (ATP) molecules, the main energy transferring molecule within cells. Figure 4.1 illustrates the important role mtDNA-encoded proteins play in cellular energy production.

Staining properties of mtDNA have indicated that many copies of mitochondrial transcription factor A (TFAM)

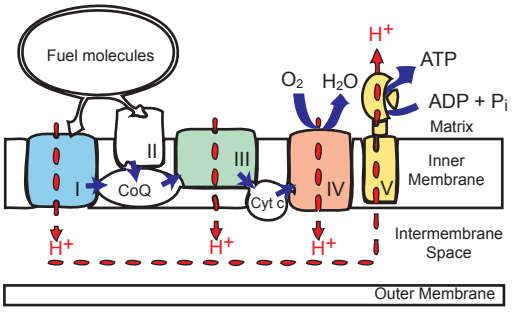
a. Human mtDNA (~16.6kb)



b. Energy generation in mitochondria



c. ATP production - the OXPHOS pathway



d. Nuclear and mtDNA subunits of the OXPHOS system

Complex	I	II	III	IV	V
Enzyme name	NADH-CoQ Reductase	Succinate-CoQ Reductase	CoQ-Cytochrome C Reductase	Cytochrome C Oxidase	ATP Synthase
Nuclear DNA Subunits	39	4	10	10	~14
Mitochondrial DNA Subunits	7	0	1	3	2

Figure 4.1 The role of mtDNA in cellular energy production

a) The 13 proteins encoded by mtDNA are subunits of the OXPHOS system, and are coloured here for complexes: complex I proteins are blue, complex III green, complex IV pink and complex V yellow. All but one of the genes (ND6) are encoded by the light strand of the mtDNA molecule. tRNA positions are shown as solid dark blue lines. The divisions between ND4 and ND4L, and the overlapping ATP6 and ATP8 are indicated by narrow black lines.

b) Carbohydrates and fatty acids are imported into the mitochondria where they are converted by ß-oxidation and the tricarboxylic acid (TCA) cycle into fuel molecules for the OXPHOS system which generates ATP.

c) The OXPHOS system receives electrons from the fuel molecules and transfers these sequentially through the complexes, with the aid of two electron shuttle molecules Coenzyme Q (CoQ) and Cytochrome c (Cyt c), to Complex IV where they are accepted by oxygen. This process generates free energy which is used to pump protons (H^+) from the mitochondrial matrix to the intermembrane space, creating an electrochemical gradient across the inner mitochondrial membrane, which is used by Complex V to drive the formation of ATP from ADP and phosphate (Pi).

d) The relative contributions of nuclear and mtDNA encoded proteins to the formation of the OXPHOS complexes.

Adapted in part by permission from Macmillan Publishers Ltd: Nature Reviews Genetics: Taylor and Turnbull 2005 (p390), copyright 2005, with additional information from the website of the Neuromuscular Disease Center, Washington University, St. Louis, MO., USA http://www.neuro.wustl.edu/neuromuscular/

surround each molecule, forming histone-like protective structures (Larsen *et al.* 2005). Within the mitochondrial matrix mtDNA molecules are present in groups termed 'nucleoids' which contain between two and 10 mtDNA molecules, each replicating independently of others in the nucleoid (Iborra *et al.* 2004, Legros *et al.* 2004). Iborra *et al.* (2004) found that nucleoids appeared to be tethered directly or indirectly through the mt membrane to kinesin and microtubules in the cellular cytoplasm. Their results suggested that the nucleoid structures are involved in the organisation of the translation machineries on both sides of the mt membrane, allowing mtDNA encoded proteins to be synthesised close to nuclear encoded proteins targeted to the OXPHOS complexes. Fluorescent imaging studies have demonstrated that nucleoids are dynamic, moving both within and between mitochondria (Garrido *et al.* 2003, Legros *et al.* 2004). The presence of multiple nucleoids within each mitochondrion, and multiple mitochondria within each cell can result in mtDNA copy numbers of several thousand per cell.

The two strands of the circular mtDNA molecule are described as 'heavy' (H) and 'light' (L) according to the different buoyant densities of each due to different guanine and thymine base compositions (Taanman 1999). The antisense sequence of most genes is encoded on the H-strand. The L-strand encodes eight tRNAs and a single protein, ND6. The reference sequence (Appendix B, Anderson *et al.* 1981, Andrews *et al.*

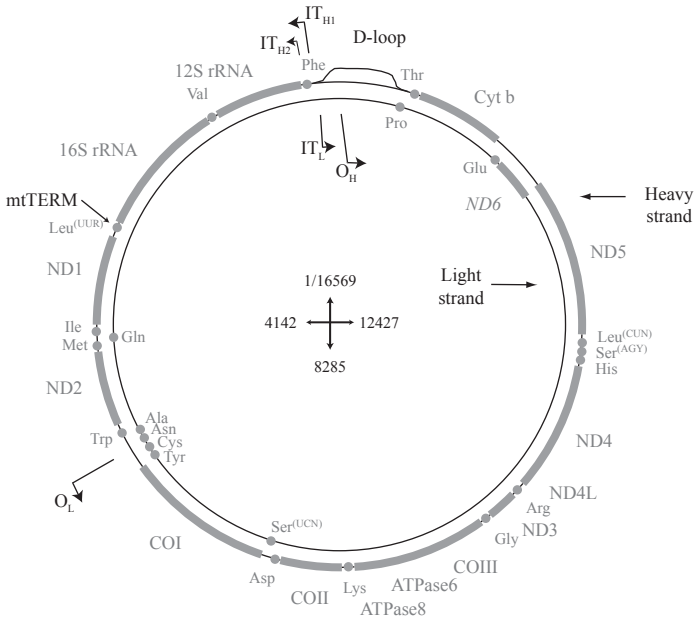


Figure 4.2 mtDNA transcription and replication features
Numbers in the centre of the diagram refer to the base numbering of the reference sequence (AC_000021.2). Protein coding and rRNA genes are represented by grey lines and tRNAs by grey circles. IT: initiation of transcription positions; H, heavy strand, L, light strand. O_H and O_L, origin of replication, heavy and light strands respectively. mtTERM proposed transcription termination site for heavy-strand transcript from IT_{H1}. Adapted from Taanman 1999:p105.

1999) represents the L-strand. The genome is very compact, with few non-coding portions apart from the ~1kb control region, and instances of overlap between genes (ATP6 and ATP8: 46 nucleotides; and ND4 and ND4L: seven nucleotides). The termination codons are in many cases generated by post-transcriptional polyadenylation of the mRNA rather than encoded in the mtDNA (Fernández-Silva *et al.* 2003). The two main non-coding regions have important functions in regulating transcription and replication: the control region contains the origins of replications for the H-strand and promoters for transcription of both strands, and the second shorter region (located between nt5730-5760) contains the origin of replication for the L-strand (Figure 4.2).

Transcription

The current model of H-strand transcription describes two origin sites - one within the control region (IT_{H1}) and a second close to the 5' end of the 12S rRNA gene (IT_{H2}), allowing for differential regulation of rRNA and mRNA transcription. Transcription from IT_{H1} ends at the 3' end of the 16S rRNA gene (termination is linked to a regulatory site - mtTERM - located immediately downstream of the 16S rRNA gene, Figure 4.2), and results in the synthesis of 12S and 16S rRNA and two tRNAs. This transcription unit is estimated to be produced at a frequency 20 times greater than the second which starts from IT_{H2} (Fernández-Silva *et al.* 2003). The mRNA transcription unit beginning at IT_{H2} encompasses all of the H-strand genes in a single polycistronic mRNA molecule. The mature mRNAs for all 12 proteins and the 12 tRNAs encoded on the H-strand are derived from processing of this second transcription unit, with the interspersed tRNA molecules believed to act as signals for the processing molecules under a model of 'tRNA punctuation' (Fernández-Silva *et al.* 2003, Taanman 1999).

The H-strand genes are transcribed in a single molecule from an initiation site in the control region (IT_L). In addition to the ND6 mRNA and eight tRNAs encoded by the H-strand, short RNA primers necessary for H-strand replication are derived from the L-strand transcription unit; linking the regulation of mtDNA replication to mtDNA transcription (Fernández-Silva *et al.* 2003). The exact locations and mechanisms for termination of the L-strand and the second H-strand transcription units have not been determined (Fernández-Silva *et al.* 2003).

The initiation points IT_{H1} and IT_L are situated within regions highly conserved in mammalian species consisting of a promoter element of 15 nucleotides surrounding the initiation site and a second element immediately upstream which includes the binding site for mitochondrial transcription factor A (TFAM: promoter elements and transcription binding sites are marked in Figure 5.2). The mitochondrial RNA polymerase requires mitochondrial transcription factor A and one of two other mitochondrial transcription factors, B1 or B2 to initiate transcription (Gaspari *et al.* 2004). The initiation point at IT_{H2} does not have the same promoter features, supporting an alternative model of a single H-strand initiation point (Taanman 1999,

Taylor and Turnbull 2005).

Mature mRNAs are translated on the mitochondrial ribosomes, using just the 22 tRNAs encoded by mtDNA: the mitochondrial code differs from the 'universal' code, allowing translation with considerably fewer tRNAs than required by nuclear-derived mRNA. Both the mitochondrial rRNAs and tRNAs are smaller than those from eukaryotic nuclear DNA, and prokaryotic cells (Fernández-Silva *et al.* 2003, Taanman 1999).

Replication and repair

mtDNA is replicated by DNA polymerase γ, a heterodimeric enzyme which is mitochondria-specific and consists of a catalytic subunit with proof-reading ability and a processivity subunit. Other proteins, including Twinkle, which has 5' to 3' DNA helicase activity, and a mitochondrial single-stranded binding protein are involved in mtDNA replication (Taylor and Turnbull 2005).

The origin of replication of the H-strand (O_H) involves several sites within the control region (Fish *et al.* 2004), while the light-strand origin (O_L) is situated in a small non-coding region between the tRNA genes for asparagine and cysteine, (nucleotide positions 5730-5760; Fernández-Silva *et al.* 2003, Figure 4.2). Until recently mtDNA replication was thought to occur under a 'strand-displacement' model where the replication of the H-strand begins first (the leading strand) and continues until the L-strand origin of replication is reached at which point the lagging L-strand replication begins (Clayton 1982). A second model of replication has been proposed following recent two-dimensional gel electrophoresis studies which supports a bidirectional, strand-coupled mechanism and this subject remains contentious (Fish *et al.* 2004, Taylor and Turnbull 2005).

The production of reactive oxygen species (ROS) as a byproduct of the OXPHOS system is believed to contribute significantly to mtDNA damage, with oxidative attack leading to base lesions, strand breaks and cross-links to other molecules (Stuart and Brown 2006). Mitochondrial DNA repair pathways differ from nuclear DNA (reviewed in Larsen *et al.* 2005): while there is evidence of a base excision repair pathway, nucleotide excision repair does not occur, and it is unclear whether a mismatch repair pathway is present. The repair of complex damage such as double-strand breaks and DNA cross-links requires a recombination repair mechanism (Stuart and Brown 2006). Homologous recombination activity has been demonstrated with mitochondrial protein extracts (Thyagarajan *et al.* 1996), and recent studies have reported evidence of recombination events in human mtDNA (D'Aurelio *et al.* 2004, Zsurka *et al.* 2005, Zsurka *et al.* 2007). An experiment using mice cells also found evidence of recombinant molecules in low frequencies (3/318 clones, Sato *et al.* 2005).

mtDNA inheritance

mtDNA has a markedly different pattern of inheritance to chromosomal DNA: the haploid sperm and oocyte

cells contribute chromosomal DNA equally at fertilisation to create the diploid zygote while the mtDNA is derived from the maternal oocyte along with all of the cytoplasmic contents of the zygote. The possibly of paternal 'leakage' of mtDNA exists as sperm cells contain mitochondria in the midpiece where they generate the energy required to power the cell's movement. Estimates of the number of mtDNA molecules present have had quite different results; from less than 10 to 700-1200 (Jansen and Burton 2004). A pathway for the destruction of paternally-derived mitochondria has been described: mitochondria entering the oocyte are directed to a 26-S proteasome by recognition of a ubiquitin tag, which is applied during spermatogenesis (Sutovsky *et al.* 2004). The ultimate fate of paternal mtDNA itself is not so well-known, although both lysosomes and autophagic vacuoles contain acid hydrolases which can degrade nucleic acids, and nuclease activity has been demonstrated in proteasomes (Sutovsky *et al.* 2004). A recent study using Japanese medaka (a small fish species) as a model organism tracked paternal mtDNA in sperm using SYBR green I staining (Nishimura *et al.* 2006). Their results suggest that the sperm mtDNA is digested within the mitochondrion shortly after fertilisation, and before the proteolytic degradation of paternal mitochondria. There is direct evidence in humans that this mechanism is not fool-proof: paternal inheritance of mtDNA has been demonstrated in a patient where the paternal haplotype was present in skeletal muscle, with the maternal haplotype dominant in all other tissues tested (Schwartz and Vissing 2002).

Any mutation arising in the mtDNA pool within an individual cell will result in the state of heteroplasmy - a mixture of two (or more) different mtDNA haplotypes within an individual. Eventually this variation will be lost as either the new or the original haplotype reaches fixation. Early research suggested this time to fixation was short, with sequence variation rapidly lost or fixed within lineages (Aquadro and Greenberg 1983), largely due to a bottleneck event occurring during the transmission of mtDNA between generations. However, the evidence for a bottleneck event has not been clear, and in a 1997 review Lightowlers *et al.* stated 'the numerous uses of 'bottleneck' by many authors are clearly inconsistent and it is our belief that this imprecision has led to confusion in the field…. there are several reports in man, mice and *Drosophila*, as well as in cell culture systems, of stable or persistent intergenerational heteroplasmy (i.e. slow segregation), results that are incompatible with a common sampling event or pattern of marked genetic drift. Paradoxically, therefore, it is not the rapid genotypic shifts that are difficult to explain, it is stable heteroplasmy that is problematic' (Lightowlers *et al.* 1997:453).

Bendall and Sykes (1995) studied the inheritance patterns of heteroplasmic length variants from the human control region, and concluded persistence of heteroplasmy was more likely than *de novo* regeneration of the variation each generation. The inheritance of heteroplasmy is common in human clinical disorders, for example hereditary optic neuropathy (LHON) and mitochondrial myopathy, encephalopathy, lactic acidosis and stroke-like episodes (MELAS) are maternally inherited, and caused by point mutations in mtDNA-encoded complex I genes. In MELAS patients the variants are present in affected individuals along with fully functional mtDNA in a heteroplasmic mix, while LHON patients can have heteroplasmic or homoplasmic

mtDNA (Taylor and Turnbull 2005). In many heteroplasmic mtDNA disorders the proportions of the mtDNA types within the body have been shown to vary between tissue types (Taylor and Turnbull 2005). A pedigree study of a family with members affected by LHON has shown persistence of heteroplasmy over six generations (Howell *et al.* 2005).

An accurate quantification of mtDNA molecules at the many stages of germ cell development is required to assess the bottleneck effect. In 1998 an influential study quantifying the number of mitochondria in human germ cells from electron micrographs concluded that during an early phase of oogenesis, in primordial germ cells (PGCs), the total number of mitochondria present was around 10 per cell (Jansen and de Boer 1998). Mitochondria at this stage were stated to be 'haploid' - each containing a single mtDNA molecule, resulting in a severe bottleneck effect. This was suggested to be a mechanism to counteract the asexual Muller's ratchet effect of the accumulation of deleterious mutations by random genetic drift (Van Blerkom *et al.* 2004; Krakauer and Mira 1999 argue the high proportion of oocytes undergoing atresia in fetal life fulfils a similar purpose).

More recently the development of real-time PCR techniques have allowed for increased precision in the estimate of mtDNA molecule numbers. Reynier *et al.* (2001) found a wide range of mtDNA copy numbers in unfertilized human oocytes collected from 43 subjects, from 20 000 to 598 000, with an average of 193 000. May-Panloup *et al.* (2007) report an even wider of variation between 11 000 and ~900 000. Current work quantifying mtDNA molecules in Chinook salmon eggs is producing higher estimates than expected, in the order of 10^9 molecules per egg (Jonci Wolf, personal communication). Cao *et al.* (2007) quantified mtDNA molecules in mouse germ-cell development stages and found in contrast to Jansen and de Boer (1998) that PGCs contain more than 100 molecules; concluding that any bottleneck effect is due to the differential segregation of nucleoid groups of mtDNA.

mtDNA recombination

The persistence of heteroplasmy through generations and the demonstration of recombination between human mtDNA molecules (D'Aurelio *et al.* 2004, Zsurka *et al.* 2005, Zsurka *et al.* 2007) suggests that mtDNA may not always follow the simple pattern of strict maternal inheritance of a single haplotype that is assumed. Recombination of multiple haplotypes occurring in germ cells could result in a phylogenetic reconstruction of following generations that may not reflect the actual history of the mtDNA lineage (a diagrammatic representation of intra-individual, or intra-lineage recombination is shown in Figure 4.3). Paternal leakage of mtDNA molecules, and subsequent recombination between maternal and paternal lineages, as seen in human muscle tissue (Schwartz and Vissing 2002, Kraytsberg *et al.* 2004) would also result in misleading mtDNA phylogenies.

The question of recombination in human mtDNA has generated a great deal of debate (for example Dowton

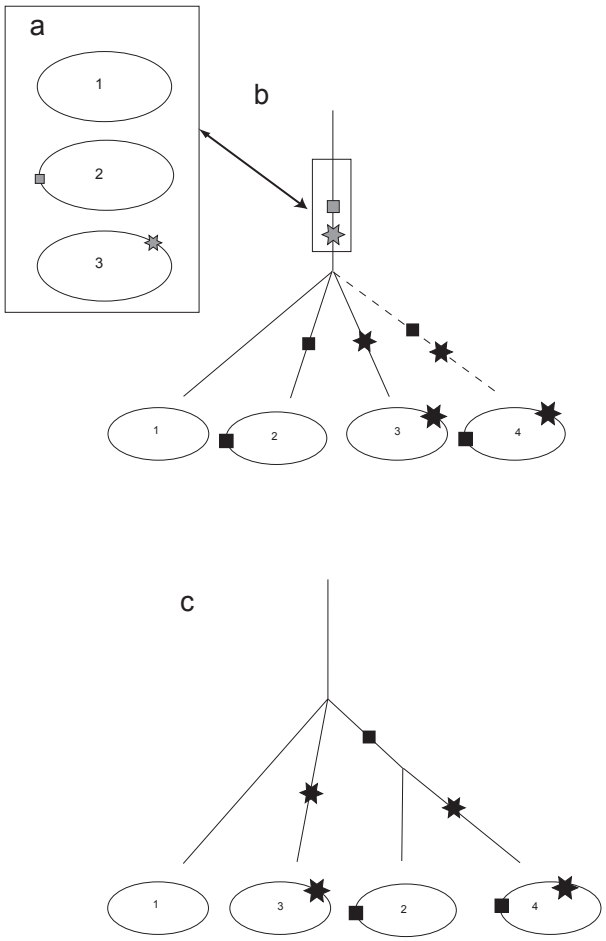


Figure 4.3 Intra-individual mtDNA recombination

Figure a) a theoretical ancestor has a heteroplasmic mix of three mtDNA haplotypes: the original inherited haplotype (1), and two new haplotypes each differing at one position from the original marked by a square (2) and a star (3). These are shown in grey to represent the three molecules are present in heteroplasmy.

Figure b) is a diagrammatic representation of possible sorting of the three ancestral haplotypes to homoplasmic descendants, with recombination. The SNPs are shown in grey when present in heteroplasmy, and black once fixed. The recombinant descendant (4) is marked with a dotted line.

Figure c) shows an example of a phylogenetic reconstruction of the four descendant haplotypes, without recombination, demonstrating the distortion of the history of the haplotypes. In this tree haplotypes (2) and (4) are assumed to have a MRCA subsequent to shared ancestry with haplotypes (1) and (3), and two independent mutations at the SNP marked by a star are required. (An equally parsimonious reconstruction would group haplotypes (3) and (4) together, and assume a repeat mutation at the square SNP).

and Campbell 2001, Macaulay *et al.* 1999, Eyre-Walker 2000, Hagelberg 2003, Hey 2000, Slate and Gemmell 2004), particularly following reports of experimental evidence in an Island Melanesian population (later corrected, Hagelberg *et al.* 1999a, 2000) and indirect evidence for recombination in primate mtDNA (Awadella *et al.* 1999, Eyre-Walker *et al.* 1999a). Other indirect analyses have found little, or limited, evidence for recombination in human mtDNA (Ingman *et al.* 2000, Elson *et al.* 2001, Herrnstadt *et al.* 2002, Piganeau and Eyre-Walker 2004). Homoplasy tests report excessive convergent changes, which would be expected if recombination was occurring (for example Eyre-Walker *et al.* 1999b, Piganeau and Eyre-Walker 2004). Whether this excess of homoplastic sites is due to recombination or extreme intra-sequence rate variation (with highly homoplastic sites often referred to as 'hypervariable'; Innan and Nordberg 2002, Stoneking 2000) is difficult at present to differentiate.

Hypervariable sites and selection

The mtDNA control region, particularly the hypervariable regions I and II, has been a preferred target for phylogenetic studies due to its high rate of variability. This rate of substitution is not consistent across all of the sequence, with some bases showing much higher rates of change. This has led to suggestions that changes

at these bases may reflect recombination events (Hagelberg 2003). Several studies have analysed patterns of change in this region (for example Malyarchuk *et al.* 2002, Malyarchuk 2004, Meyer *et al.* 1999, Stoneking 2000). New mutations have been observed to occur preferentially at hypervariable sites (Forster *et al.* 2002, Stoneking 2000) and the current consensus is that these positions are more mutable than others. Malyarchuk *et al.* (2002) propose that a transient misalignment dislocation of DNA polymerase γ due to the strand context (monotonous runs of nucleotides) can explain the substitution rate variability within the control region. Recent studies of entire mtDNA sequences have also identified positions within the coding region which require more steps to fit a tree than expected (Kivisild *et al.* 2006, Pierson *et al.* 2006).

The mt genome plays a critical role in cell function, with the many known medical conditions due to mtDNA mutations providing clear examples of the deleterious effects changes can have on an individual's fitness (for example see Gemmell *et al.* 2004, Taylor and Turnbull 2005). An interesting example of effects on cell conditions linked to mtDNA variability has been shown by Kazuno *et al.* (2006), who analysed transmitochondrial hybrid cells to detect ribosomal and non-synonymous protein-coding changes in the context of functional significance using a fluorescent calcium indicator. They found that cells carrying mtDNA molecules with 8701A/10398A variants (these are macrohaplogroup N-defining) had different mitochondrial matrix pH and calcium levels, relative to individuals with 8701G/10398G haplotypes (macrohaplogroups L and M). Transitions at these positions result in substitutions between alanine and threonine amino acid residues in ATP6 and ND3 proteins.

Several studies have tested human mtDNA data sets for evidence of selection (Mishmar *et al.* 2003, Moilanen and Majamaa 2003, Elson *et al.* 2004, Ruiz-Pesini *et al.* 2004, Kivisild *et al.* 2006). Mishmar *et al.* (2003) suggested from their results of analyses on a data set of 104 sequences that climate-driven selection may be acting on lineages in northern latitudes. Ruiz-Pesini *et al.* (2004) suggested haplotypes common in Artic regions contained variants which resulted in a lowered coupling efficiency of the OXPHOS system, resulting in greater heat and reduced ATP production. Amo and Brand (2007) have tested this hypothesis empirically by kinetic analysis of cybrid cells, finding the 'Arctic' haplotypes have similar or greater coupling efficiency than the 'tropical' types defined by Ruiz-Pesini *et al.* (2004).

A 'mother's curse' effect has been proposed whereby certain mtDNA haplotypes may have an adverse effect on sperm motility (Gemmell *et al.* 2004, Ruiz-Pesini *et al.* 2000); however a recent study of whole mtDNA sequences from males with reduced sperm motility did not find a correlation between haplotype and sperm failure (Pereira *et al.* 2007). Moilanen and Majamaa (2003) found evidence for haplogroup specific differences in the intensity of selection against particular regions of the mitochondrial genome, and suggested the effects of new mutations could be dependent on the phylogenetic background on which they arose. However Kivisild *et al.* (2006) did not observe any significant regional or climactic differences in the rate of non-synonymous changes for haplogroups, or evidence of directional selection.

4.2 Assembly of the whole mtDNA data sets and haplogroup assignment

Entire and coding-region only human mtDNA sequences available from NCBI databases up to May 2006 were downloaded and manually aligned using SE-AL (version 2.0a11, Rambaut 1996) into a large data set which included the coding region from 4 great ape sequences. Table 4.1 summarises the accession, publication and geographic details of these sequences. 1844 of the 2631 sequences are complete mtDNA genomes and 787 coding-region only sequences. This alignment was constructed incrementally from 2003 onwards, with sequences added as they became available. The majority of sequences are the first version uploaded to the databases, and it is possible that revised versions now exist. Where revised versions are known to have replaced earlier sequences it is noted in Table 4.1.

The alignment of 2631 sequences was exported from SE-AL as a FASTA file, which allowed it to be opened in MacClade (version 4.06 Madison and Madison 2003, Sinauer Associates Inc., Massachusetts); when opening from NEXUS format the maximum number of sequences allowable is 1500). The 16686 characters of the alignment were numbered for the Mitomap rCRS sequence (accession AC_000021.2) prefaced with an 'n', using the 'fill' utility in MacClade. Where gaps were added to the reference sequence (which is 16569 bases long) these are named for the preceding character with an alphabetical suffix; for example the alignment character number 598 corresponds to n573a, which is an insertion relative to the reference sequence directly following nucleotide 573 in the Mitomap rCRS. The file was saved from MacClade in NEXUS format as 'globalmtDNA.nex' (Appendix F.4.1.1). NEXUS format character labels are supported by MacClade and PAUP* (version 4.0b10, Swofford 2003), and were used in two main ways in this study: firstly when tracing characters on trees produced by MacClade, and secondly, for excluding single bases or portions of the alignment in PAUP*. For example the control region can be excluded from the alignments in PAUP* using this command: 'excl n1-n577; excl n16024-n16569'. A list of alignment character numbers with the corresponding rCRS base numbering was extracted from the NEXUS file and is recorded in a Microsoft® Excel workbook ('globalmtdata set.xls': Appendix F4.1.7).

MacClade was used to find identical sequences (when resolution of missing data or ambiguities could make sequences identical) and found 280 of the 2631 sequences belonged to 116 haplotypes (details listed in 'globalmtdata set.xls'). The majority of these haplotypes (86) were shared by just two individuals, with a maximum of six individuals with a common haplotype (in two instances, both containing individuals only from the Moilanen *et al.* (2003) Finnish data set). A separate file 'globalmtDNAhaps.nex' (Appendix F4.1.2) contains only the haplotype sequences (number of taxa = 2468), with the final taxon in alphanumerical order representing each of the shared haplotypes.

Two other subsets of the global data set contain only the complete mt genome sequences, and are named

Table 4.1 Global data set details

Accession marked with an asterisk are prefaced with a 'z' in data sets to enable rapid sorting

Authors	Year	Geographic notes	n	Accession numbers
Ingman et al	2000	Global	53	AF346963-AF347015
Maca-Meyer et al	2001	Global	33	AF381981-AF382013
Herrnstadt et al (and Howell et al 2004)	2002	U.S. (global) – most coding only, L0-L3 entire mtDNA	597	Numbered xxxHerrn
Ingman and Gyllensten	2003	Global; focus on Australia and PNG	52	AY289051-AY289102
Kong et al	2003	China	48	AY255133-AY255180
Maca-Meyer et al	2003	N/R/U haplogroup	11	AY275527-AY275537
Mishmar et al	2003	Global	48	AY195745-AY195792
Moilanen et al	2003	Finland (includes sequences described in Finnila et al '01. **NOTE:** revised versions for some sequences available)	192	AY339402-AY339593
Achilli et al	2004	N/R/H haplogroup	62	AY738940-AY739001
Coble et al	2004	R/R/HV, N/R/JT, N/R/U/K	241	AY495090-AY495330
Palanichamy et al	2004	India – N only	75	AY713976-AY714050
Tanaka et al	2004	Japan	672	AP008249-AP008920
Achilli et al	2005	N/R/U haplogroup	39	AY882379-AY882417
Friedlaender et al	2005	New Britain, New Ireland	3	AY956412-AY956414
Macaulay et al	2005 (version 2)	Malaysia	15	AY963572-AY963586
Merriwether et al	2005	Near Oceania	14	DQ137398-DQ137411
Thangaraj et al (2005)	2006 (version 2)	Andaman and Nicobar Islands	15	AY950286-AY950300
Trejaut et al	2005	Taiwan	8	AJ842744-AJ842751
Rajkumar et al	2005	India	23	DQ246811-DQ246833
Starikovskaya et al	2005 (version 2)	Siberia	20	AY519484-AY519497 AY570524-AY570526 AY615360-AY615361
Behar et al	2006	Ashkenazi Jews	30	DQ301789-DQ301818
Kivisild et al	2006	Global - coding region only	277	DQ112686-DQ112962
Li. S et al (unpubl.)	2006	China, Tibet, Mongolia	5	DQ418488 DQ462232-DQ462234 DQ437577
Pierson et al	2006	Oceania, Taiwan	20	DQ372868-DQ372887
Sun et al	2006	India	56	AY922253-AY922308
Thangaraj et al (unpub.)	2006	India (assumed)	9	DQ408672-DQ408680
van Holst Pellekaan	2006 (version 2)	Australian	8	DQ404440-DQ404447
Anderson et al (1980), Andrews et al (1999)	2007 (version 2)	Europe revised 'Cambridge Reference Sequence'	1	AC_000021.2*
Arnason et al	1996	Gorilla; coding only	1	X93347*
Arnason et al	1996	Chimpanzee (Pan troglodytes); coding	1	X93335*
Horai et al	1995	Chimpanzee (Pan paniscus); coding	1	NC_001644*
Xu and Arnason	1996	Orangutan; coding only	1	NC_002083*

Table 4.2 Data set file details (Digital Appendix F4.1)

File name	Details	No. of taxa	File size
globalmtDNA.nex	All sequences in nexus format	2631	42.2MB
globalmtDNAhaps.nex	Identical sequences removed, haplotypes only	2468	39.6MB
globalmtDNAcomplete.nex	Complete mtDNA sequences only, nexus format	1844	29.6MB
globalmtDNAhapscomplete.nex	Complete mtDNA haplotype sequences only, nexus format	1736	27.9MB
globalmtDNAhaps1.meg	Haplotypes only, MEGA format annotated for genes. Group categories: coding-only, entire sequences, Ingman 200 sequences, Kivisild 2006 sequences	2468	52.6MB
globalmtDNAhaps2.meg	Haplotypes only, coding-region only. MEGA format annotated for genes and for haplogroup affiliation (33 groups)	2468	49.1MB
globalmtdataset.xls	Excel workbook with complete haplotype sequences, Genbank and publication data for all sequences, comparative character numbering and shared sequences information	1736	27.9MB

'globalmtDNAcomplete.nex' and 'globalmtDNAcompletehaps.nex' (Appendices F4.1.3 and F4.1.3). The 'globalmtDNAhaps.nex' alignment was edited using BBEdit Lite or TextWrangler (Bare Bones Software, Bedford, MA, USA, available from www.barebones.com) creating two input files for MEGA (Kumar *et al.* 2004, version 3.1). MEGA allows sections of a sequence to be annotated, for example as protein-coding genes, and subsequent subsets of the characters to be analysed independently. The mtDNA control region and genes of both heavy and light strands were recorded in the files 'globalmtDNAhaps1.meg' (Appendix F4.1.5) and 'globalmtDNAhaps2.meg' (Appendix F4.1.6), with areas of overlap treated consecutively. For example the ATP8 gene begins within the ATP6 gene, but as each character can only be within one assigned domain or gene the overlapping region is recorded as ATP6. The other regions of overlap involve single nucleotides in tRNA sequences. Genes encoded on the light strand (eight tRNAs and ND6) are marked as ':light in the MEGA descriptions. The first MEGA data set has 16683 characters, in contrast to the 16686 characters of the 'globalmtDNAhaps.nex' file: three single-base insertions in protein-coding genes were removed as they caused frame-shift mutations. All three were singleton (private) insertions and occurred in sequences 170Herrn (n3312a, ND1), DQ246818 (n5436a, ND2) and AY956413 (12355a, ND5).

The two MEGA files differ in the groups to which the 2468 sequences are assigned, and the second file excludes the control region for all sequences. The first file has a simple division to five categories: 'coding only', 'entire sequence', 'Ingman 2000 sequences' and 'Kivisild 2006 sequences' and 'apes'. In the second file the sequences are divided to 33 groups by haplogroup designation, including the 'apes' group of 4 sequences (the method of haplogroup assignment is described below). These groupings, in combination with the sequence data annotations, allow rapid independent analysis within MEGA of subsets of the mt

genome data set by haplogroup affiliation as well as by sequence area. Table 4.2 summarises the details of the NEXUS and MEGA format data sets.

A Microsoft® Excel workbook ('globalmtdata set.xls', Appendix F4.1.7) was used to compile publication, haplogroup and geographic and/or ethnographic details for all sequences. Details were extracted from Genbank accession files and the original publications using text editing functions within BBEdit Lite. When haplogroup information was retrieved from published figures find, copy and paste functions were used wherever possible to minimise copying errors. For example, the original sequence identifier used in the publication and described in the Genbank accession file was copied from the 'globalmtdata set.xls' file and pasted into a search of the PDF publication file. Haplogroup information from the published figure was then entered into the 'globalmtdata set.xls' file.

The original haplogroup designations of Howell *et al.* (2004), Kivisild *et al.* (2006), Kong *et al.* (2003), Maca-Meyer *et al.* (2001), Mishmar *et al.* (2003), Palanichamy *et al.* (2004), Sun *et al.* (2006) and Trejaut *et al.* (2005) collected to the 'globalmtdata set.xls' file from the Genbank format accessions and/or publications were used to assign broad haplogroups to all of the sequences in the global data set, following the construction of a low-resolution consensus tree from the coding region of all of the human sequences in the 'globalmtDNA.nex' data set. This tree was obtained by an heuristic search using PAUP* version 4.0b10 (Swofford 2003) for most parsimonious trees which was stopped after 18 hours (running on an Intel ® Pentium ® 4 CPU 3.00GHz, 2.00GB of RAM) when 10 391 trees had been saved with parsimony score 6476 (number of taxa 2627, 2134 parsimony informative coding region characters; gaps were treated as missing data).

Table 4.3 Global data set haplogroup details

Level 1	n	% of total dataset
L0	39	1.48
L1	58	2.21
L2	78	2.97
L3	2448	93.19
L4	2	0.08
L5	2	0.08

Level 2	n	% of total dataset
L1/b	26	0.99
L1/c	32	1.22

Level 2	n	% of total dataset
L3/(bd)/f	44	1.67
L3/ei	30	1.14
L3/h	4	0.15
L3/N	1638	62.35
L3/M	732	27.86

Level 3	n	% of total dataset
N/A	86	3.27
N/N1	40	1.52
N/W	55	2.09
N/X	20	0.76
N/O	4	0.15
N/R	1358	51.69

Level 3	n	% of total dataset
M/D	300	11.42
M/G	67	2.55
M/M7	92	3.50
M/M8	76	2.89
M/Q	22	0.84

Level 4	n	% of total dataset
R/B4a	46	1.75
R/B4bd	48	1.83
R/B5	38	1.45
R/HV	552	21.01
R/JT	225	8.56
R/P	18	0.69
R/R5	11	0.42
R/R9	56	2.13
R/U	301	11.46

Level 5	n	% of total dataset
JT/J	107	4.07
JT/T	118	4.49

A 75% majority-rule consensus of these trees was exported in PDF format from PAUP* and opened in Adobe® Illustrator® CS2 (version 12.0.1, © 1987-2005 Adobe Systems Incorporated and its licensors). The original haplogroup designations from the studies listed above were appended to the accession numbers using the search and replace functions in Adobe® Illustrator® CS2 and Microsoft® Excel. Broad haplogroups were then assigned conservatively to the tree according to these labelled sequences: in instances where the monophyly of the named haplogroups was not supported by the consensus tree these branches were assigned to the corresponding higher level haplogroup (Appendix D4.1). For example, sequences originally assigned to haplogroups L3/M/M1 (DQ112926 and DQ112933) share a common branch from the M vertex with four others including two sequences labelled M11 (AF381984) and M12 (AF381996): these are all named 'M' in the consensus tree (see Appendix D4.1, branch directly above M7bc/c labelled branch).

Blocks of sequences were copied by haplogroup from the consensus tree file and haplogroup identifiers added using BBEdit Lite before importing the information to the 'globalmtdata set.xls' workbook. Up to five haplogroup levels were assigned for each sequence and these are arranged in separate columns enabling searching and sorting functions to be carried out. Table 4.3 summarises the haplogroup names and number of sequences assigned to each level. A large proportion of the sequences belong to the 'Out-of-Africa' L3 haplogroup - 2448 of the total 2627 (93%), and of these 1358 (52% of the data set) belong to L3/N/R haplogroups. The over-represented haplogroups N/R/HV, N/R/JT and N/R/U haplogroups are common in European populations.

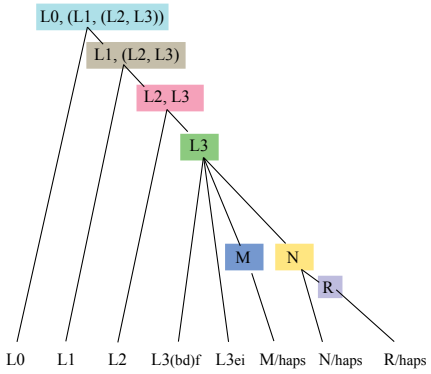
4.3 Variation in the global data set

Average pairwise distances between haplogroups

The maximum absolute human pairwise distance in the global data set is 108 base differences (0.65%) across the entire sequence. For the coding region only the maximum distance is 90 (0.59%, calculated from the globalmtDNAhapscomplete.nex and globalmtDNAhaps.nex data sets). As a simple contrast, the average absolute distance over the 15317 sites in the coding region alignment between the human and the two chimpanzee sequences is 1283 (8.4%; and between human and the single gorilla and orang-utan sequences 1590 (10.4%) and 2168 (14.15%) respectively). Figure 4.4 summarises the average distances between representative haplogroups (those containing >15 haplotypes) calculated from the coding region only (using MEGA, Kumar *et al.* 2004, version 3.1).

The summary divergence values for the L3 haplogroups suggest that variation in the coding region has accumulated at a steady rate (Figure 4.4). The M and N vertices differ from the L3 vertex by three and four coding region polymorphisms respectively, and the R vertex by just one change from N (Torroni *et al.* 2006), and their descendants share an average of ~12.5 differences to their respective common ancestors. The range of average distances between haplogroups within L3, with the exception of M, is greater than those for the

n	Haplogroup	L0	L1	L2	L3	L3(bd)f	L3ei	L3/M	M/D	M/G	M/M7	M/M8	M/MQ	L3/N	N/A	N/N1	N/W	N/X	N/R	N/R/HV	N/R/U	N/R/JT	N/R/B4	N/R/B5	N/R/P	N/R/R9
36	L0																									
54	L1	70.0																								
76	L2	70.2	65.2																							
2294	L3	70.2	64.2	44.9																						
43	L3(bd)f	68.8	62.4	43.6																						
28	L3ei	66.3	59.2	39.6		24.3																				
694	L3/M	70.2	64.4	43.5		28.8	24.7																			
281	M/D	70.6	64.5	43.5		28.8	24.7																			
64	M/G	71.2	64.3	44.0		29.8	25.5		23.9																	
92	M/M7	72.1	66.1	45.8		30.9	26.8		24.7	25.9																
70	M/M8	71.5	66.3	44.6		30.8	26.5		24.6	25.6	26.9															
20	M/MQ	71.6	66.0	45.5		30.7	26.5		24.5	25.8	26.3	26.6														
1525	L3/N	70.4	64.3	45.7		30.7	26.7	30.8	30.5	31.7	32.5	32.6	32.5													
81	N/A	70.3	64.1	45.9		30.8	26.6	30.8	30.6	31.9	32.8	32.7	32.6													
40	N/N1	71.0	64.6	47.1		31.6	27.7	29.5	30.3	31.6	32.6	32.2	32.3		25.8											
43	N/W	71.6	65.2	47.9		33.2	29.3	33.2	33.4	32.7	35.0	35.1	34.1		25.7	26.8										
17	N/X	68.5	61.1	43.6		27.7	23.9	27.6	28.5	29.7	30.8	30.0	34.1		20.7	21.7	23.3									
1266	N/R	70.6	64.6	45.9		30.9	26.9	31.0	30.7	31.9	32.7	32.8	32.8		23.6	25.0	26.0	20.2								
504	N/R/HV	67.3	61.1	42.6		27.7	23.5	27.5	27.1	28.7	29.8	29.5	29.5		19.7	22.6	22.4	17.5								
276	N/R/U	74.3	68.3	49.6		34.8	30.6	34.7	34.6	35.6	36.8	36.5	36.5		27.5	29.8	30.1	25.2		22.3						
210	N/R/JT	72.7	66.5	47.8		32.9	28.6	32.8	32.3	33.3	34.7	34.9	34.7		26.1	27.1	27.8	22.6		20.8	28.0					
93	N/R/B4	70.0	63.9	44.0		30.2	26.2	30.4	30.3	31.1	31.9	32.1	31.8		22.4	23.5	25.0	19.3		17.2	25.0	23.6				
36	N/R/B5	74.2	67.6	49.0		34.4	30.4	34.2	34.6	33.7	36.7	34.3	36.5		28.7	29.8	29.6	25.4		23.6	30.6	28.1	25.7			
18	N/R/P	72.1	66.0	47.7		32.8	28.7	32.8	32.6	33.9	35.0	34.6	34.8		25.0	27.6	27.6	22.5		19.8	27.5	25.0	22.4	28.8		
56	N/R/R9	74.9	69.0	50.3		35.2	31.4	35.5	35.4	36.6	37.5	37.4	37.4		27.5	30.3	30.3	25.3		22.5	30.2	28.7	25.2	31.5	27.7	



	# of pairwise comparisons	Min	Max	Max-Min	Average	SD
MRCA (L0, (L1, (L2, L3)))	24	66.3	74.9	8.6	70.9	2.0
MRCA (L1, (L2, L3))	23	59.2	69.0	9.8	64.7	2.3
MRCA (L2, L3)	22	39.6	50.3	10.7	45.6	2.5
MRCA L3	117	23.5	37.5	14.0	31.5	3.2
MRCA M	10	23.9	26.9	3.0	25.5	1.0
MRCA N	38	17.5	30.3	12.8	25.2	3.3
MRCA R	21	17.2	31.5	14.3	25.4	3.8

Figure 4.4 Absolute pairwise distances (coding region) between groups in the global data set
Top: Average absolute distances between groups, coloured for their relatedness within the human mtDNA tree. Bottom left: The major nested clades of the human mtDNA tree. Bottom right: A summary of the maximum (max), minimum (min), range (max-min), average and standard deviation (SD) distances for major points in the mtDNA tree.

deeper comparisons, but this may be a factor of the much larger number of haplotypes present in L3.

Variability by mtDNA region

The percentage of variable sites, V(sites), and variable amino acids V(a.a.) where applicable, for different parts of the mt genome over the global data set are listed in Table 4.4. Counts were generated by MEGA, (Kumar *et al.* 2004, version 3.1) from the 'globalmtDNAhaps1.meg' data set of 2464 human haplotypes with the exception of the values for the control region which are taken from the 1736 complete sequences in the data set. A separate file was constructed from the reverse complement of the ND6 gene to examine its amino acid variability. The percentage of variable sites and amino acids in each region has been calculated for all

Table 4.4 Variable sites summary

Abbreviations: V, variable. non-S, non-singleton. a.a., amino acid. The 'independent' category groups all non-coding bases from the coding region.

Region	N (sites)	% V (sites)	% V non-S (sites)	% V(a.a.)	% V non-S (a.a.)
Control region	1161	39	29		
All tRNAs	1516	15	8		
12S rRNA	960	15	9		
16S rRNA	1572	11	7		
ND1	957	26	16	20	11
ND2	1044	25	16	21	14
COI	1542	20	13	11	5
COII	684	24	15	16	9
ATP8	207	39	24	52	35
ATP6	681	38	23	45	28
COIII	782	25	14	22	12
ND3	348	22	14	10	9
ND4L	297	18	11	6	4
ND4	1373	21	14	11	7
ND5	1812	24	16	18	12
ND6:light	525	26	16	23	14
Cyt.b	1137	29	18	28	18
Independent	131	32	18		

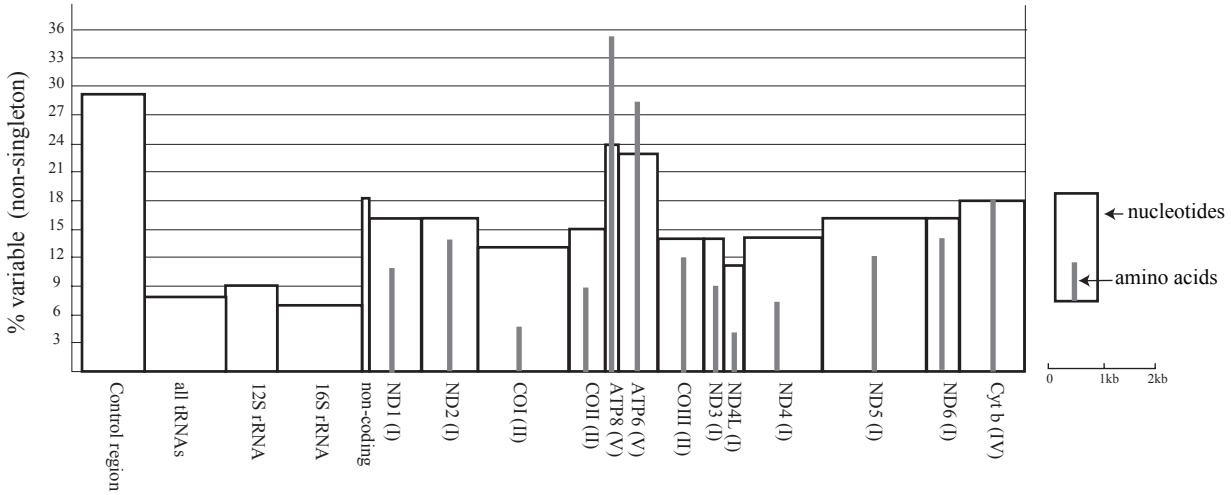


Figure 4.5 Sequence and amino acid variation by region

The percentage of non-singleton variable bases and amino acids are charted for gene regions, with OXPHOS complexes indicated in brackets. The width of columns for nucleotides (in white) is proportional to the gene's length.

Table 4.5 Variability in tRNA genes

Gene	N (sites)	% V	% V non-S
tRNA-Ala	69	17	9
tRNA-Arg	65	12	8
tRNA-Asn	73	5	3
tRNA-Asp	69	12	9
tRNA-Cys	67	28	18
tRNA-Gln	72	21	15
tRNA-Glu	69	13	7
tRNA-Gly	68	18	9
tRNA-His	69	14	7
tRNA-Ile	70	9	4
tRNA-Leu	75	7	4
tRNA-Leu	72	14	8
tRNA-Lys	70	13	7
tRNA-Met	68	7	4
tRNA-Phe	73	15	5
tRNA-Pro	75	13	5
tRNA-Ser	73	12	7
tRNA-Ser	60	20	8
tRNA-Thr	67	42	28
tRNA-Trp	68	15	7
tRNA-Tyr	67	9	7
tRNA-Val	69	14	3

changes (% V) and for non-singleton changes (% V non S): excluding the changes present only in single haplotypes (singletons) results in a more conservative estimate of variation but helps to reduce any ‘noise’ that may be present due to sequencing errors. The average number of non-singleton changes per region are charted in Figure 4.5.

The control region has a high proportion of variable sites (39% overall, 29% when singleton polymorphisms are excluded), consistent with the expectation that this non-coding region accumulates variability rapidly compared to other parts of the mt genome. The RNA genes have much lower levels of variable positions; 15% or less, while the protein-coding genes show quite a broad range of variability, from 18% overall for ND4L to 39%, the same as the control region, for ATP8 (including singleton polymorphisms). ATP6 and ATP8 are the only mtDNA encoded proteins to contribute to ATP Synthase, Complex V of the OXPHOS system (Figure 4.1). Overall, approximately half of the amino acids of ATP8 (35/69), and almost half of the ATP6 amino acids (102/227) are variable. The Cytochrome *b* gene has the most sequence variability following the ATP6 and ATP8 genes, with 29% nucleotide and 28% amino acid variability.

The percentage of variable sites for the tRNA genes when grouped together is 15% (Table 4.4). Table 4.5 lists the variation within each gene independently, revealing a wide range of variation (from 5% to 42% of

informative variable sites). The greatest polymorphism is found in the $tRNA^{Thr}$, which is situated between the Cytochrome *b* gene and $tRNA^{Pro}$ (28/67 sites in total are variable). Kivisild *et al.* (2006) also reported an excess of changes in this gene compared to other tRNAs, and examined whether the observed changes seen in their data set would be likely to affect the function of the molecule. As none of the changes fell within the 100% conserved region for the mammalian consensus sequence they concluded that it was likely it had not lost its function.

4.4 Tests of selection

Three tests of selection were carried out on protein coding subsets of the global data set: Tajima's *D* test, the McDonald-Kreitman test and calculations of non-synonymous to synonymous substitution ratios (ka/ks). Tajima's test examines the difference between two measures of sequence diversity; the number of segregating sites and the average number of nucleotide differences estimated from pairwise comparison (Tajima 1989). This relationship is used to test the neutral mutation hypothesis (Kimura 1968). Subsets by gene of the global data set were tested using Arlequin (version 3.01, Excoffier *et al.* 2005), with 1000 simulated samples to estimate p-values. The Arlequin infiles were prepared through the export function in DnaSP (version 4.10.9, Rozas *et al.* 2003). The original 2468 haplotypes (including four great ape sequences) were reduced to 2435 to enable import into DnaSP which will not accept sequences with ambiguity codes. This is a considerably larger dataset than those previously analysed (for example Kivisild *et al.* 2006 n=277, Mishmar *et al.* 2003 n=104). The 33 excluded haplotypes were identified using SplitsTree (version 4.6, Huson and Bryant 2006) and are listed in the 'globalmtdataset' Excel file (Appendix F4.1.7).

The Tajima *D* test results are shown in Table 4.6. The values obtained are all significantly negative ($p<0.001$, with the exception of ND3 where $p<0.01$), indicating an excess of low-frequency polymorphisms, consistent with selective effects (for example a selective sweep event), a bottleneck effect and/or a recent expansion in population size (Kreitman 2000).

The McDonald-Kreitman test for selection compares the ratio of amino acid replacement substitutions (ka) to synonymous substitutions (ks) within and between species, as under neutral evolution this ratio should be the same (McDonald and Kreitman 1991). This test was done for each gene in the global data set using DnaSP, with the two chimpanzee entire mt genome sequences in the alignment (*Pan paniscus* NC_001644 and *Pan troglodytes* X93335) identified as the second species (Table 4.6). For two genes, Cyt *b* and *ND2*, larger chimpanzee data sets were available from public databases, and the results for these are listed alongside the results for the test with the n=2 chimpanzee data set (details of additional sequences are given in the caption to Table 4.6). The results from the COIII and Cyt *b* data sets differ significantly from neutral expectations ($p<0.05$); however when the larger chimpanzee data set is included the Cyt *b* p-value is not significant at the 95% level.

	Tajima's test		McDonald-Kreitman test			
			Pan sp. n=2		*Pan* sp. n>2	
Gene	***D***	***P***	**NI**	***P***	**NI**	***P***
ATP6	-2.61	0.0000	1.86	0.25		
ATP8	-2.48	0.0001	1.38	1.00		
COI	-2.63	0.0000	2.48	0.17		
COII	-2.60	0.0000	0.89	1.00		
COIII	-2.56	0.0000	5.64	0.01		
CytB	-2.53	0.0000	2.91	0.02	3.12	0.09
ND1	-2.66	0.0000	1.05	1.00		
ND2	-2.56	0.0000	3.18	0.07	3.80	0.07
ND3	-2.23	0.0012	0.86	1.00		
ND4	-2.54	0.0000	0.48	0.06		
ND4L	-2.37	0.0003	1.29	1.00		
ND5	-2.60	0.0000	1.08	0.88		
ND6	-2.52	0.0000	1.24	1.00		

Table 4.6 Tests of selection by gene: Tajima's D and McDonald-Kreitman results
The p-values for the Tajima's *D* test are approximated to the beta distribution by simulation of 1000 samples within Arlequin (version 3.01, Excoffier *et al.* 2005), and for the McDonald-Kreitman test are Fisher's two-tailed exact test values calculated within DnaSP (Rozas *et al.* 2003). NI: neutrality index.
The Cyt b data set with additional chimpanzee sequences consisted of a portion of the gene only (254/379 codons, nt14747-nt15509) and added 16 haplotypes from sequences AY585833-AY585844 (Morin, P.A., unpublished, direct submission 30/3/04, Molecular Ecology Laboratory, Southwest Fisheries Science Center, La Jolla, CA 92037, USA) and EF660764-EF660819 (Lorenz,J.G. *et al.*, unpublished, direct submission 8/6/07 Molecular Biology Laboratory, Coriell Institute for Medical Research, Camden, NJ 08103, USA). Sequences with ambiguous bases or shorter in length were excluded. All sequences are . *troglodytes*. The ND2 data set with additional chimpanzee haplotypes included sequences AF440167-AF440192 (Stone *et al.* 2002); from both *P. paniscus* and *P. troglodytes*, excluding sequences with ambiguities, increasing the number of chimpanzee haplotypes to 19.

Pairwise ratios of non-synonymous to synonymous substitutions (Ka/Ks) in the protein coding genes for the global dataset (n=2431 human sequences, sequences with ambiguities excluded as described above) were also calculated using DnaSP. The large output files were analysed using Microsoft® Excel and the statistical package R (R version 2.5.1 (27/06/07) copyright 2007: The R Foundation for Statistical Computing ISBN 3-900051-07-0). To avoid dividing by zero a constant representing a single synonymous substitution was added to the denominator, so that the ratio calculated was Ka/(Ks + constant), following the method used in analyses by Kivisild *et al.* (2006) and Mishmar *et al.* (2003). The average of these values for each gene are shown in Figure 4.6, and compared with results reported by Kivisild *et al.* (2006, n=277), and Mishmar *et al.* (2003; n=104, note that the values are estimated from their Figure 2: p174, to the nearest 0.05).

The large increase in sample size from the earlier studies to the analysis of 2431 sequences here appears to affect the results from the Ka/(Ks+constant) calculations for several genes; Cyt *b* and *ND3* (Figure 4.6)

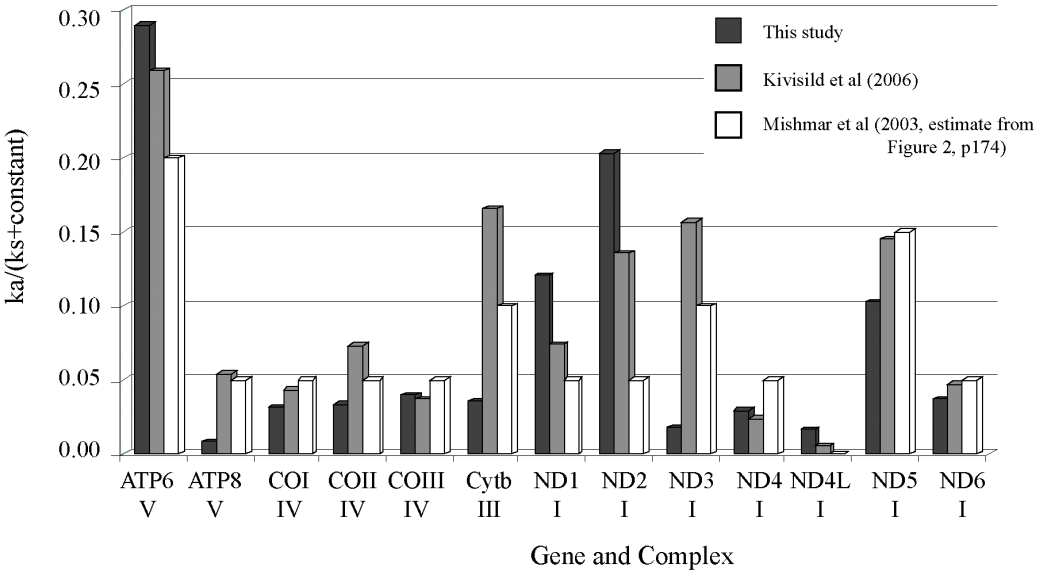


Figure 4.6 Chart of average ka/ks+constant values

in particular. However, the skewed distribution of the global data set towards high frequencies of L3/N haplotypes (Table 4.3), and the use of average values in order to make comparison between this study and those of Kivisild *et al.* (2006) and Mishmar *et al.* (2003) may have contributed to the differences between the studies seen in Figure 4.6.

A striking result from the Ka/(Ks+constant) analyses are the high ratios seen in all three studies for the *ATP6* gene compared to the other protein coding genes (from this study the average is 0.29, and a value of 0.5 represents equal numbers of non-synonymous and synonymous substitutions). High levels of amino acid diversity within *ATP6* were apparent from the descriptions of variability within the genes obtained (Figure 4.5), with both *ATP6* and *ATP8* genes showing a high (>45%, Table 4.4) percentage of variable amino acids. The *ATP6* and *ATP8* genes encode components of Complex V of the OXPHOS system, and it is interesting that the *ATP8* gene, which shows a higher level of amino acid diversity than *ATP6,* has much lower Ka/(Ks+constant) values.

4.5 Testing for recombination

The phi (Φ, pairwise homoplasy index) statistic (Bruen *et al.* 2006) implemented in SplitsTree 4.2 was used to test a number of subsets of the global data set for evidence of recombination. This test is based on the principle of refined incompatibility and has been shown to perform well in both simulations and on empirical data for closely and distantly related samples. The data sets tested included the known recombinant sequences

Table 4.7 Phi test for recombination

Values significant at p<0.05 are boxed

Data set	n	p-value		
		entire sequence	**coding only**	
Oceania-127	138	0.40855	n.d.	
B4a	48	0.18002	n.d.	
B5a	12	0.01943	0.05272	
M7bc with M22	47	0.00953	0.11742	
M27 with M28	15	0.64729	n.d.	
P with R21	24	0.75270	n.d.	
Q with M29	27	0.67025	n.d.	
N/R/HV	254	0.89497	0.99498	
N not R	191	0.97846	n.d	
Australian set Oc-133	35	0.71954	n.d.	
M/D	261	0.60756	n.d	
all L3	1644	0.99878	n.d	
all global mt haps	1736	0.99960	n.d	
		nt11206-nt1490		
Kraytsberg (2004)	19	0.00317		
		entire sequence	**coding only**	**control region**
Random_1	30	0.10477	0.12417	0.88320
Random_2	30	0.07862	0.11729	0.15879
Random_3	30	0.56065	0.92751	0.11505
Random_4	30	0.00933	0.28853	0.00308
Random_5	30	0.88943	0.98682	0.30408
Random_6	30	0.45286	0.99761	0.37112
Random_7	30	0.01857	0.16803	0.02613
Random_8	30	0.35491	0.64176	0.59089
Random_9	30	0.39979	0.70993	0.32318
Random_10	30	0.51900	0.84918	0.21486
Random_11	30	0.41851	0.78724	0.06960
Random_12	30	0.43192	0.39865	0.68779
Random_13	30	0.01377	0.13877	0.02069
Random_14	30	0.02181	0.12925	0.57897
Random_15	30	0.32987	0.47060	0.38559
Random_16	30	0.03994	0.26875	0.00256
Random_17	30	0.07541	0.01625	0.73069
Random_18	30	0.50852	0.83402	0.49332
Random_19	30	0.61163	0.47277	0.73224
Random_20	30	0.98894	0.99000	0.91871
Random_21	30	0.28395	0.80621	0.08041
Random_22	30	0.15517	0.50428	0.16245
Random_23	30	0.10694	0.40000	0.43959
Random_24	30	0.02208	0.08965	0.35448
Random_25	30	0.18922	0.46319	0.15618
Random_26	30	0.02293	0.17904	0.02818
Random_27	30	0.05850	0.65449	0.00421
Random_28	30	0.22050	0.21788	0.56326
Random_29	30	0.01977	0.22004	0.00199
Random_30	30	0.00156	0.04486	0.01302

from an individual with paternal mtDNA inheritance (Kraytsberg *et al.* 2004), encompassing a ~7kb portion of the mt genome from nt11206-1490. Eight data sets from the analyses in Chapters Two and Three, three large subsets of the global dataset, and the entire global haplotypes data set were also tested over the entire sequence length. In addition to these data sets, thirty random sets of 30 haplotypes were drawn from the global dataset and tested over the entire sequence length, and for coding region and control region subsets of the genome.

The results of the Φ test are shown in Table 4.7. The Kraytsberg *et al.* (2004) data set showed significant evidence for recombination ($p<0.0032$), as did two of the haplogroup data sets analysed in Chapter Three; N/R/B5a ($p<0.0194$) and M7bc with M22 ($p<0.00953$). The Kraytsberg data set contains 19 haplotypes; the rCRS, maternal (N/R/HV) and paternal (N/R/U) sequences, and the 16 unique recombinant clones sequenced. The sequence length is 6882 nucleotides, from nt11206 to nt1490; there are 20 parsimony informative sites, and an heuristic search found 264 most parsimonious trees with scores of 32. Ten of the 20 characters required more than one step to fit the trees. The consensus network of the most parsimonious trees is shown in Figure 4.7.

The N/R/B5a data set was the smallest tested, with just 12 haplotypes. The phylogeny constructed from the MMS analysis of this haplogroup is shown in Figure 3.6. Recurrent mutations were seen at three of 17 parsimony informative positions (nt210, nt16266 and nt16183), all of which are within the control region. When the control region was excluded from the sequences the test values were not significant.

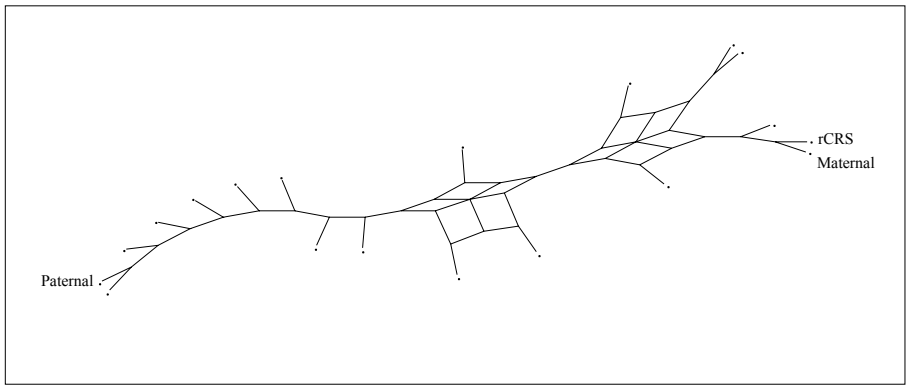


Figure 4.7 Consensus network of most parsimonious trees, Kraytsberg et al (2004) recombinants
Consensus network of 264 trees found for the Kraytsberg *et al.* (2004) recombinant data set. Unlabelled sequences are recombinants between the maternal and paternal haplotypes.

The M7bc data set (Figure 3.5) consisted of 47 haplotypes from the M/M7bc haplogroup and the single sequence named as M22 (Macaulay *et al.* 2005).The consensus network generated from the most parsimonious trees found was complex, (Figure 3.5a) and 19 of the 55 parsimony informative characters required two or more steps to fit the 5040 trees found. Twelve of these recurrent polymorphisms occurred within the control region. Of the seven coding region polymorphisms four are shared between sequences AP008278 and the M7b subhaplogroup. This is unusual, and while it may be strong evidence of a recombination event, it may simply be a result of errors during or following sequencing of the AP008278 sample (for example missed transitions at nt6455 and nt9824, Figure 3.5b). Attempts to contact the authors (Tanaka *et al.* 2004) to confirm these polymorphisms have not been successful to date.

Nine of the 30 random data sets also showed evidence of recombination using the Φ test when the entire sequence was tested (Table 4.7). The results were not consistent when the data sets were reduced to coding and control region subparts, although the final data set did give a positive signal of recombination for both subsets as well as the entire sequence. The p-values for the significance of four of these tests (three from the control region only, and one, data set 30, for the entire sequence) were smaller than that from the Kraytsberg data set (p=0.00317, Table 4.7). A graphic illustration of variation within one of the random data sets is given in Figure 4.8, which shows the consensus network for the 'Random_30' dataset constructed from the 266 most parsimonious trees found (there were 94 parsimony informative characters, and the heuristic search score =177). This data set had the lowest p-value of all tested using the Φ statistic, including the Kraytsberg *et al.* (2004) data set. Incompatibilities between the trees found appear to be concentrated around the branching to the N and N/R haplogroups present, with a smaller area of conflict at the divergence of M8, M7 and M/D haplogroups (Figure 4.8). The sites requiring more than one step to fit the trees (56/94 parsimony informative characters) are listed in the caption to Figure 4.8.

4.5 Discussion

The use of mtDNA to address issues in human prehistory has accelerated rapidly in the past decade, and the common assumption of a lack of recombination and a neutral, or nearly neutral mutation rate underlies the interpretations of these phylogenies and the date estimates obtained from them. High instances of homoplasy in the phylogenies constructed from Oceanic haplogroups in the Chapters Two and Three, and the importance of date estimates, particularly at the N/R/B4a1a1 vertex, to current models of Oceanic prehistory, prompted me to explore the processes involved in fixation of mtDNA polymorphisms in human populations in this and the following chapter.

The assembly of the global data set described in this chapter has continued throughout the course of this project, and as more sequences become available the power of these data to address issues in intraspecific mtDNA evolution will increase even further. At present the L3/N haplotypes are over-represented in the data

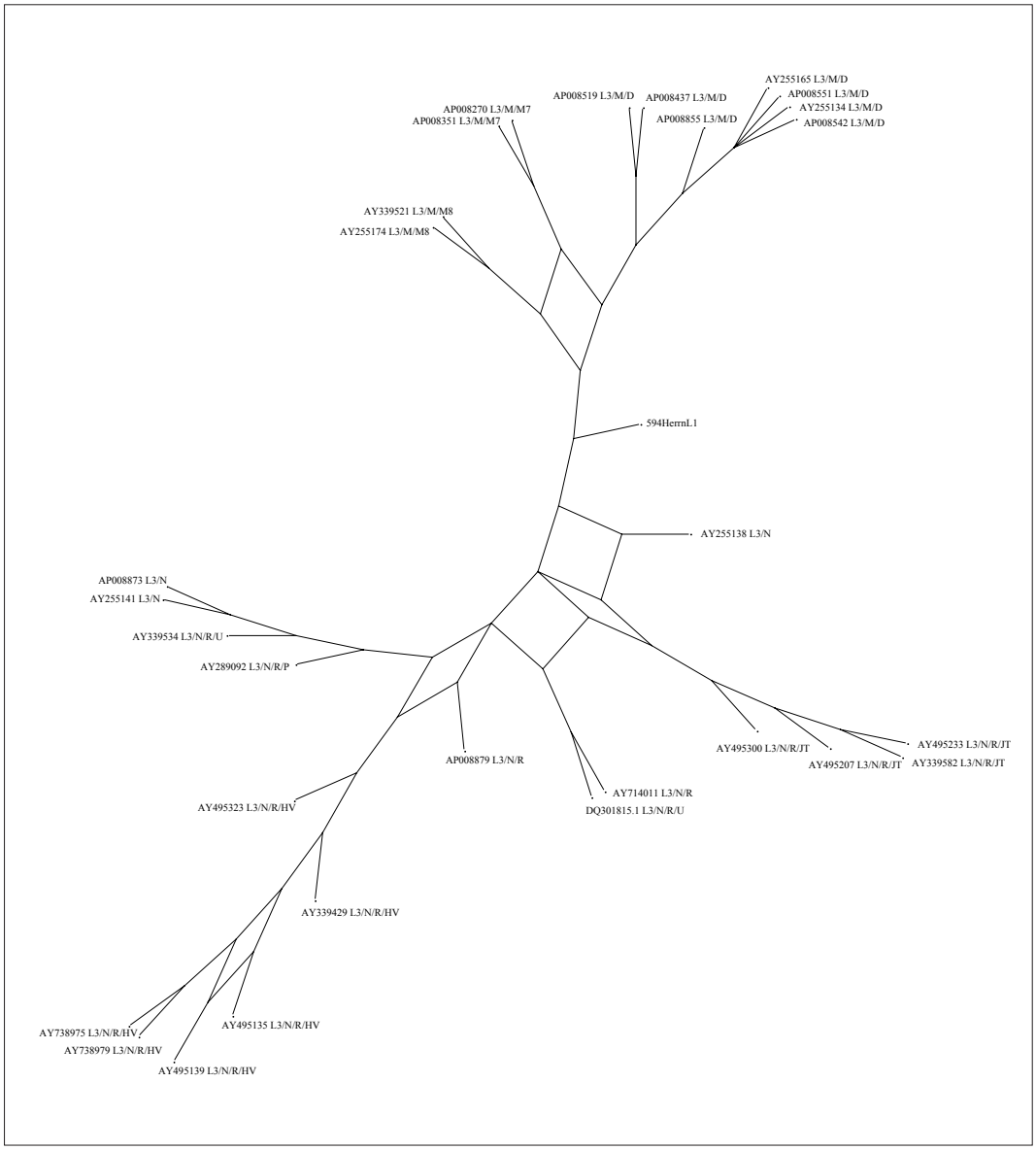


Figure 4.8 Consensus network of most parsimonious trees, Random_30 dataset entire mtDNA
Consensus network of 266 trees, number of parsimony informative characters=94, heuristic search score=177. The accession numbers are followed by haplogroup designations obtained from the 'globalmtdataset.xls' workbook (Digital Appendix F4.1.7). The phi test p-value for this data set is 0.00156.

Sites requiring more than one step in the trees are: n146, (4 steps); n150, (3); n151, (3); n152, (6.4); n182, (2); n195, (3); n489, (2); n629, (2); n709, (2); n1442, (3); n1811, (2); n2706, (2); n3010, (3.6); n3397, (2); n3666, (2); n4216, (2); n4386, (2); n5417, (2); n5460, (2); n7389, (2); n8701, (2); n10398, (2); n10685, (2); n10810, (2); n11467, (2); n11719, (2); n11969, (2); n12007, (2); n12308, (2); n12372, (2); n12705, (2.2); n13105, (4); n13708, (3); n14178, (2); n14798, (2); n15067, (2); n15607, (2); n16093, (2); n16126, (1.8); n16129, (4); n16172, (3); n16183, (2); n16189, (6); n16223, (3); n16224, (2); n16257, (2); n16261, (2); n16266, (2); n16274, (2); n16278, (2); n16292, (2); n16294, (2); n16298, (3); n16362, (2); n16390, (2); n16519, (5).

set (Table 4.3), but future analyses may help to adjust the balance; for example a recent study (Gonder *et al.* 2007) describes 62 new African sequences.

Selective advantages and disadvantages of different mtDNA haplotypes have been suggested as contributing factors for several human diseases, and the aging process (for example Coskun *et al.* 2004, Tanaka *et al.* 2004; reviewed by Taylor and Turnbull 2005), however the pathogenicity of many identified variants has been disputed from phylogenetic, and statistical arguments (for example Bandelt *et al.* 2007, McFarland *et al.* 2004). Mishmar *et al.* (2003), and Ruiz-Pesini *et al.* (2004) proposed a role for climatic adaptation in the evolution of present human mtDNA variation, but were criticised for their definitions of climatic zones and comparisons between haplogroups of differing age (comparisons were made for Ka/Ks ratios between 'older' African L haplogroups and 'younger' Eurasian L3 haplogroups). This study took the approach of examining each gene independently for evidence of selection, using the entire global data set, without division to geographic subsets as undertaken by Mishmar *et al.* (2003) and Ruiz-Pesini *et al.* (2004).

The Tajima's D values obtained for each gene from the global data set were all significantly negative, indicating an excess of low-frequency polymorphisms, which could be interpreted as evidence for a selective sweep event, or recent bottleneck and/or population expansion events. As Kreitman (2000) has noted, human population history does not readily fit the requirements of an equilibrium-neutral model, with several alternative explanations (for example dramatic changes in population size) available to explain departures from equilibrium-model predictions.

The results of the McDonald-Kreitman test identified significant differences in the divergence between humans and chimpanzee lineages in the COIII and Cyt *b* genes, while the Ka/Ks+constant calculations highlighted high ratios of non-synonymous changes in the ATP6, ND1, ND2 and ND5 genes relative to the other protein coding genes. Comparisons between the results of this study and those of Mishmar *et al.* (2003) and Kivisild *et al.* (2006) are limited by the differences in sampling between the earlier data sets and the L3/N dominated global data set.

Subsequent to this analysis, Ingman and Gyllensten (2007) have addressed earlier limitations of the studies of selection to different climates, generating a geographically balanced data set by combining 61 new mt genome sequences with 104 existing sequences. They analysed the ratios of synonymous to non-synonymous polymorphisms within protein coding regions, with reference to the positions of functional domains for each gene. On average the non-synonymous sites outside of the core functional regions of each protein were found to evolve at almost five times the rate of those within the functional domain. The genes showing the highest levels of non-synonymous polymorphisms, ATP6, ND3 and Cyt *b*, have the smallest core domains relative to gene length, explaining a large part of the differences in Ka/Ks ratios between genes. Comparisons between

the four regions analysed (North Asia, South Asia, Europe and Africa) did not support the hypothesis of mtDNA adaptation to cold climates, and the authors argue that the variation present can be best explained by relaxed purifying selection at some positions in combination with the effects of random drift (Ingman and Gyllensten 2007).

The effect of purifying selection has also been suggested as a key factor in explaining differences in mutation rates obtained from pedigree and phylogenetic rates of human mtDNA mutation (Howell *et al.* 2003, Howell *et al.* 2004). An excess of non-synonymous mutations in the terminal branches of phylogenies has been reported when compared to deeper branches (Moilanen *et al.* 2003, Moilanen and Majamaa 2003, Kivisild *et al.* 2006), consistent with a gradual effect of purifying selection acting to remove slightly deleterious variants from the population.

The phylogenetic approach to assessing selection used by Kivisild *et al.* (2005) on a data set of 277 sequences could be extended in future work to incorporate all available mt sequences. This would provide a more powerful means of examining the histories of human mtDNA molecules for evidence of the action of selection, through analysis of the likely impact of changes to protein-coding and RNA genes, and the distribution of these changes in different branches of the tree and different regions of the world.

The positive results for recombination from the phi test (Bruen *et al.* 2006) of subsets of the global data sets are intriguing, with several (almost one third) of the 30 random data sets generated showing incompatibility values consistent with recombination events. The occurrence of recombination between maternal and paternal lineages has been demonstrated (Kraytsberg *et al.* 2004), and the recombinant sequences (between N/R/HV and N/R/U haplotypes) gave a strong signal of recombination using the phi test (Table 4.7).

It is not likely that existing methods for detecting recombination events will recognise instances of intra-lineage recombination events (Figure 4.3), as the molecules involved differ only at a very small number of sites, and the recombination events are relatively rare. For example, recombinant sequences were observed at a low frequency (approximately 0.7% of the total mtDNA) in the muscle tissue of the individual with mixed maternal and paternal inheritance of mtDNA (Kraytsberg *et al.* 2004). In addition, this model of polymorphism generation is dependent upon the maintenance of heteroplasmy through generations, rather than rapid shifts between haplotypes as new mutations arise and are either eliminated from the pool of mtDNA molecules or replace the earlier ‘parental’ haplotype.

A potential example of intralineage recombination was seen in samples from the N/R/B4a haplogroup described in Chapters Three. The combination of character states at nt6905 and nt16247, and the observation

of heteroplasmy at nt16247 (Chapter Six) fits a pattern of persistent heteroplasmy and recombination, and one which would be relatively straightforward to explore further with additional N/R/B4a1a1 samples.

Paternal leakage of mtDNA and subsequent recombination, and intra-lineage recombination will result in misleading phylogenies, and distort estimates of divergences from ancestral vertices. The phi test indicates several data sets tested show unexpectedly high levels of incompatibility, seen in phylogenies as instances of homoplasy - recurrent or parallel mutations. As discussed in the recombination section of the introduction to this chapter, distinguishing between the effects of ‘hypervariability’, or extreme rate variation, and recombination in the generation of homoplasy is difficult. In the following chapter a phylogenetic approach is taken to investigate the incidence and location of ‘hypervariable’ bases in mtDNA, focusing on the control region.

5. Phylogenetic Analysis of Homoplasy in the Global Data Set

In the previous chapter aspects of mtDNA structure, function and inheritance were reviewed, and the assembly of a large data set of human mt sequences described. Variation within this dataset was examined for evidence of selective and recombination effects, in light of the high levels of homoplasy identified in earlier chapters examining haplogroups from the Oceanic region. The analyses described in this chapter investigate the occurrence of homoplasy using the large global data set and taking a phylogenetic approach. One specific aim was to assign a series of weights to the control region HVR-I characters which would facilitate the phylogenetic analysis of the large number of HVR-I sequences available from Oceania.

To investigate the rates of change at positions in the HVR-I, sets of minimal trees were constructed from the coding region of mt genomes using the MMS parsimony approach, and the control region characters 'mapped' onto the trees to measure their compatibility, or number of steps required, to fit the coding region trees. A data set size of 75 taxa (the analysis is referred to as the '75-taxa' analysis) was selected as this gave high success rates when searching for minimal parsimony scores using the MMS approach. As the control region characters do not contribute to the construction of the trees, this provides an independent means of assessing their relative rates of change. Coding region characters were also assessed, highlighting 'hypervariable' bases outside of the control region.

The results from the '75-taxa' analysis were used in the second stage of this analysis. Two weight sets for the HVR-I were determined from the '75-taxa' results and tested by evaluating their power relative to the unweighted HVR-I characters at resolving haplogroup categories in sample sets of the global data set. This analysis; 'coding vs. HVR-I', sampled five haplotypes from three of 18 haplogroups to produce 5000 data sets of fifteen taxa, and examined the first most parsimonious trees found using the different character subsets (coding, HVR-I unweighted, and the HVR-I characters using the two weight sets). The power of the different HVR-I subsets at attaining the same level of haplogroup differentiation as the coding subset was assessed.

5.1 Methods used for 75-taxa analysis

Construction of the random data sets

An Excel® master workbook was used to construct 200 data sets and record details of the analysis. PAUP* (version 4.0b10, Swofford 2003) was used to reduce the 'globalmtDNAcompletehaps.nex' data set (Appendix F4.1.2) described in Chapter Four to only the parsimonious characters present (n=2012), excluding gapped sites. The

'show character status' command (full details, hiding excluded characters), outputted to the log file a table with columns: character, type, status, weight and states. The character column contained the number in the globalmtDNAcompletehaps.nex data set of the character, and in brackets the Mitomap rCRS number assigned in MacClade (version 4.06; Sinaueur Associates, Inc.). BBEdit Lite (version 6.1.2 Bare Bones Software, MA, USA) was used to edit this list into a format suitable for import into Excel® where these two numbers and a third identifier the number of the characters in the parsimony informative only data set - were stored. The states shown at each parsimony informative site were recorded in the Excel® workbook.

The parsimony informative only data set was saved out from PAUP* and imported into the Excel® workbook, alongside the haplogroup information for each sequence obtained from the global consensus tree (Chapter Four). As such a large proportion of the sequences available belong to the L3 macrohaplogroup the data set was ordered by haplogroup and split into two sections: non-L3 sequences (n=92) and L3 (n=1644). Twenty-five taxa were selected from the non-L3 group, and 50 from the L3 group to make subsets of 75 taxa. This was done by inserting a column alongside the data (non-L3 and L3 in two different worksheets) which generated random numbers (the '=RAND() command), and then copying and pasting these values into a second column. This column was used to sort the accession names and sequences and copy the top 25 or 50 of these to a new worksheet using an Excel® macro. The 75 sequences were then pasted to BBEdit Lite, and edited into NEXUS format for PAUP* input.

PAUP* and MMS analysis

Heuristic parsimony searches were carried out on the 200 data sets, and followed by MMS analysis (Holland *et al.* 2005). Four NEXUS format infiles were prepared each containing 50 data sets with PAUP* command blocks (details are provided in Appendix C5.1). Control region characters were excluded, and the remaining characters reduced to those which were parsimony informative. A PHYLIP format file was exported at this point for later input to the MMS (Holland *et al.* 2005). An heuristic search was run, and trees found saved to a PHYLIP file. All of the characters were then restored and the parsimony scores for each character on the first tree found were exported to a text file. The character scores which varied between trees were listed in the log file.

The parsimony scores for each character in the first tree found were imported into the Excel® worksheet and inserted alongside the columns containing character number identification according to the original 'globalmtDNAcompletehaps.nex' data set and the Mitomap rCRS. The characters whose number of steps varied across the trees found were analysed in a separate file after extraction from the log file and editing for import to Excel® using BBEdit Lite. In several cases this required the division of the list of number of steps

per tree into subsets for import, as the number of trees found exceeded the maximum number of rows in an Excel® worksheet (65536). The number of steps per character was averaged over all most parsimonious trees, and this average replaced the score for that character previously recorded in the Excel® master file from the first most parsimonious tree. Details from the log file for each data set (number of parsimonious characters, the parsimony score and number of trees found) were extracted and saved to the Excel® workbook.

The PHYLIP format data sets were analysed using the MMS program in batch files. The score found by heuristic search for each data set was inserted as an 'upper' boundary condition for the MMS. The first 100 data sets underwent six sets of MMS analysis (parameters are recorded in Appendix C5.2), with data sets removed from the analysis as they were proved minimal. The parsimony scores of a total of 65 of the first 100 data sets were found to be minimal using this approach. The second set of 100 75-taxon data sets underwent three cycles of MMS analysis using the same parameters, after which 49 parsimony scores were found to be optimal, resulting in a total of 114 minimal data sets for the character analysis.

5.2 Results of the 75-taxa analysis

The results of the 75-taxa analysis are summarized in Appendix Table E5.1, which lists details for each of the 2012 parsimony informative characters in the data set. The majority of the parsimony informative characters (93%) have two states, and characters from the control region make up 16% of the total (318/2012). A disproportionate amount of the characters showing more than two states (50/131 of the characters with three states present, and 11/14 of the characters with all states present) are from the control region.

148 bases were parsimony informative in all of the 114 75-taxa data sets that were proved minimal using the MMS in all data sets. All of the 2012 informative characters in the entire globalmtDNAcompletehaps.nex data set were informative in at least one of the 114 data sets (two bases were informative in just one data set, 11 in two, and 19 in three). The average number of data sets overall that each of the 2012 character was present in was 39 (34% of the total of 114). The average number of steps required to fit the trees in which a character was informative in ranged from the base level of 1, where the character was compatible with the tree, to 19 (for nt16519).

Homoplasy in the coding region

While the main goal of this analysis was to explore the relative rates of mutation among bases within the control region (which did not contribute to the tree reconstructions), it also revealed instances of

Table 5.1 'Hotspots' in the coding region

Base number	States	Average steps	Pars inform. (% datasets)	Gene	Codon position	Amino acid details	Surrounding sequence
709	AG	6.6	100	12S rRNA	-	-	AAGCATCCCC G TTCCAGTGAG
962	CT	3.91	100	12S rRNA	-	-	ATCACCCCCT C CCCAATAAAG
1438	AG	2.39	100	12S rRNA	-	-	GCAGTAAACT A AGAGTAGAGT
1598	AG	1.52	83	12S rRNA	-	-	GCACTTGGAC G AACCAGAGTG
1719	AG	1.83	87	16S rRNA	-	-	GACAACCTTA G CCAAACCATT
2225	AC	1.5	1.75	16S rRNA	-	-	CACCCACTAC C TAAAAAATCC
2352	CT	1.6	100	16S rRNA	-	-	TGCGTCAGAT T AAAACACTGA
3010	AG	3.88	100	16S rRNA	-	-	AGGACATCCA G ATGGTGCAGC
4655	AG	1.51	82	ND2	3	synonymous	ATTTCCTCAC G CAAGCAACCG
5046	AG	1.58	100	ND2	1	valine(G)/ isoleucine(A)	AATAATAGCA G TTCTACCGTA
5147	AG	2.56	95	ND2	3	synonymous	CCAGCACCAC G ACCCTACTAC
5231	AG	1.76	99	ND2	3	synonymous	TAGGAGGCCT G CCCCCGCTAA
5237	AG	1.93	90	ND2	3	synonymous	GCCTGCCCCC G CTAACCGGCT
5460	AG	3.75	100	ND2	1	alanine(G)/threonine(A)	CACACTCATC G CCCTTACCAC
6221	ACT	2.08	94	COI	3	synonymous	GACTCTTACC T CCCTCTCTCC
7055	AG	1.83	100	COI	3	synonymous	TATCAATAGG A GCTGTATTTG
7867	CT	1.5	100	COII	3	synonymous	ACGATCCCTC C CTTACCATCA
8251	AG	2.74	98	COII	3	synonymous	TTGAAATAGG G CCCGTATTTA
8584	AG	1.53	96	ATP6	1	alanine(G)/threonine(A)	CCTACCCGCC G CAGTACTGAT
8701	AG	1.59	100	ATP6	1	threonine(A)/alanine(G)	ACAAATGATA A CCATACACAA
8790	AG	1.69	92	ATP6	3	synonymous	TCGGACTCCT G CCTCACTCAT
9377	AG	1.67	82	COIII	3	synonymous	TATACCAATG A TGGCGCGATG
9824	ACT	2.12	98	COIII	3	synonymous	TCCACGGACT T CACGTCATTA
10398	AG	4.18	100	ND3	1	threonine(A)/alanine(G)	ATTAGACTGA A CCGAATTGGT
10915	CT	1.6	100	ND4	3	synonymous	TATTTAGCTG T TCCCCAACCT
11299	CT	1.67	94	ND4	3	synonymous	TACTACTCAC T CTCACTGCCC
11914	AG	4.65	100	ND4	3	synonymous	TAGTAACCAC G TTCTCCTGAT
11944	CT	1.5	100	ND4	3	synonymous	TACATATTTA C CACAACACAA
12007	AG	2.23	100	ND4	3	synonymous	CAACACAATG G GGCTCACTCA
12172	AG	1.72	83	$tRNA^{his}$	-	-	CAGATTGTGA A TCTGACAACA
12372	AG	1.65	99	ND5	3	synonymous	CCCTAACCCT G ACTTCCCTAA
12501	AG	1.67	81	ND5	3	synonymous	CAATATTCAT G TGCCTAGACC
12705	ACT	1.75	100	ND5	3	isoleucine(C/T)/methionine(A)	ATCTACTCAT C TTCCTAATTA
13105	AG	2.13	100	ND5	1	isoleucine(A)/valine(G)	TGTAGCAGGA A TCTTCTTACT
13590	AG	1.97	100	ND5	3	synonymous	CTACCTCCCT G ACAAGCGCCT
13708	AG	2.87	98	ND5	1	alanine(G)/threonine(A)	TAAACGCCTG G CAGCCGGAAG
13928	ACGT	2.94	97	ND5	2	serine(G)/asparagine(A) / threonine(C)/isoleucine(T)	TTCTACCCTA G CATCACACAC
14470	ACT	1.54	77	ND6	3	synonymous	CTGTAGTATA T CCAAAGACAA
14766	CT	1.61	100	Cyt b	2	isoleucine(T)/threonine(C)	ATACGCAAAA C TAACCCCCTA
14798	CT	1.82	99	Cyt b	1	phenylalanine(T)/leucine(C)	TAACCACTCA T TCATCGACCT
15043	AG	1.74	100	Cyt b	3	synonymous	TACACATCGG G CGAGGCCTAT
15236	AG	1.63	80	Cyt b	1	isoleucine(A)/valine(G)	AGTTCAATGA A TCTGAGGAGG
15244	AG	1.78	100	Cyt b	3	synonymous	GAATCTGAGG A GGCTACTCAG
15301	AG	2.5	100	Cyt b	3	synonymous	ACTTCATCTT G CCCTTCATTA
15607	AG	1.51	93	Cyt b	3	synonymous	TCCCTAACAA A CTAGGAGGCG
15784	CT	2.09	100	Cyt b	3	synonymous	TAAGCTACCC T TTTACCATCA
15924	AG	2.56	96	$tRNA^{thr}$	-	-	AGTCTTGTAA A CCGGAGATGA

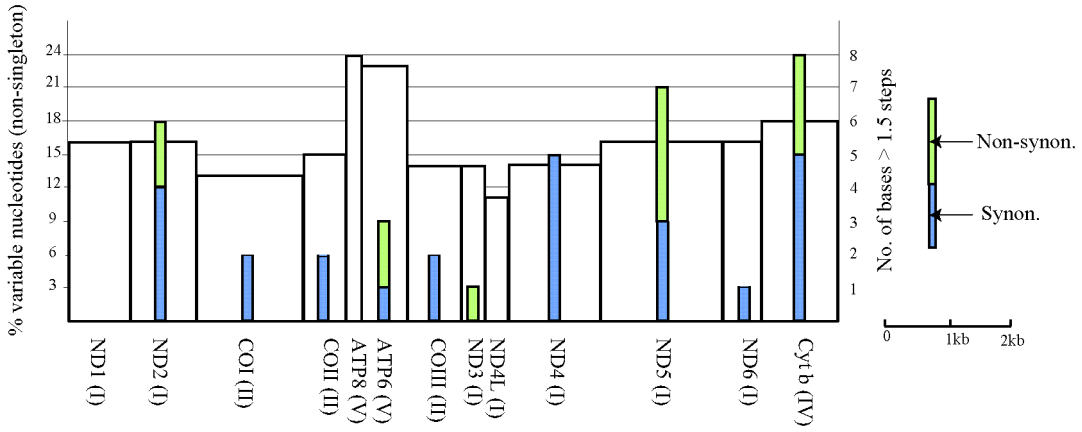


Figure 5.1 Distribution of homoplasious bases in protein-coding genes
The number of base positions requiring an average of more than 1.5 steps in the 75-taxa data set analysis are shown in shaded bars (number of steps is on the right y-axis). This information is set on the background of unshaded columns representing the percentage of informative nucleotides in each gene from the entire globalmtDNAcompletehaps.nex' data set (percent values are labelled on the left y-axis). The unshaded column width is proportional to the length of the gene in kilobases. The OXPHOS complex each protein contributes to is shown in parentheses.

'hypervariable' coding region bases. 47 bases in the coding region had an average of more than 1.5 steps in the 114 data sets. Of these, 8 were in rRNA genes (4 each within 12S rRNA and 16S rRNA), two were in tRNA genes and the remaining 37 were in the protein coding genes (Table 5.1, also marked on the annotated reference sequence, Appendix C).

Six of the 47 bases identified required an average of more than three steps to fit the trees. Three of these, nt709 (6.6 steps), nt962 (3.91 steps), and nt3010 (3.88 steps) are within the ribosomal RNA genes and involved transitions only. The three other highly homoplasious bases are sited within ND genes: nt5460 (3.75 steps, ND2), nt10398 (4.18 steps, ND3), and nt11914 (4.65 steps, ND4). The first two of these transitions cause amino acid substitutions between alanine and threonine; the third is a synonymous change.

Four of the changes in Table 5.1 occur within, or directly adjacent to poly-C sequences (nt962, nt5231, nt5237 and nt6221) and this may reflect an increased chance of mutation during replication. Of the 41 sites with two states undergoing transitions, 32 (78%) were between adenine and guanine bases. Kivisild *et al.* (2006) found a similar trend towards purine substitutions in the recurrent changes seen on a tree of 277 mtDNA sequences.

The distribution of the 37 synonymous and non-synonymous changes in the protein-coding genes is shown in Figure 5.1. The gene with the highest number of recurrent changes is Cyt b (eight), with seven and six changes occurring respectively in the ND5 and ND2 genes. Twelve of the 37 changes requiring an average of more than 1.5 steps occurring in the protein-coding genes result in amino acid substitutions (Table 5.1). The most common amino acid substitutions were between alanine and threonine (five instances), and valine and isoleucine (three instances). This pattern of change was also reported by Moilanen and Majaama (2003) and Kivisild *et al.* (2006). Three other bases with relatively high rates of change caused amino acid replacements between isoleucine and methionine, isoleucine and threonine, and phenylalanine and leucine amino acids. At nt13928, a second codon position in the ND5 gene, all four nucleotides were observed, with each resulting in a different amino acid: serine (G), asparagine (A), threonine (C) and isoleucine (T).

Three of the five characters characterising the N macrohaplogroup (transitions to nt8701A, nt9540T, nt10398A, nt10873T and nt15301T: Appendix D3.5) showed high parsimony scores in the analysis. The nt8701 (average step rate 1.59) and nt10398 (average step rate 4.18) transitions occur in the ATP6 and ND3 genes and both result in codon changes from threonine to alanine. The third change, at nt15301 in the Cyt *b* gene is synonymous, and required an average of 2.5 steps on the 75-taxa trees. The single coding region change along the N/R branch, at nt12705 (Appendix D3.5) also appears in the set of 47 bases requiring more than 1.5 steps (Table 5.1), and causes an amino acid change between isoleucine and methionine in the ND5 gene. The remaining two changes defining N; 10873T (ND4) and 9540T (COIII) are synonymous and required 1.1 and 1.0 average steps respectively.

Of the four base changes on the branch leading to the M macrohaplogroup (transitions to nt489C, nt10400T, nt14783C, nt15043A: Appendix D3.6), one (nt15043, a synonymous change in the Cyt *b* gene, 1.74 steps) required an average of more than 1.5 steps. The other two changes in protein coding genes, 10400T (ND3) and 14783C (Cyt *b*) are synonymous and were fully compatible with the trees (averaging 1.00 steps). The fourth change at nt489C occurs in the non-coding control region, and had an average parsimony score of 1.9.

The relatively high rates of change seen at the bases defining the N and R macrohaplogroups may simply reflect the over-representation of these types within the data set (Chapter 4), and the position of the changes in the tree: these sites were parsimony informative in all 114 of the data sets assessed. However, the M-defining sites were also parsimony informative in all data sets yet do not show the same pattern as some of the N-defining sites. For example, the M-defining nt10400T transition, requiring an average of 1 step, occurs in the third position of the same codon as the non-synonymous N-defining nt10398 transition which

requires an average of 4.18 steps. This appears conspicuously high, and may be an indication of selective pressure acting on this variant. As described in Chapter Four, Kazuno *et al.* (2006) have reported differences in mitochondrial matrix pH and calcium levels associated with the nt10398A polymorphism. Studies have also reported correlations between the state at nt10398A and various health factors; for example a reduced risk of Parkinson disease, (van der Walt *et al.* 2003), increased risk of bipolar disorder (Kato *et al.* 2001), and longevity (Niemi *et al.* 2005), however see Saxena *et al.* (2006) for a comprehensive review of the issues surrounding mtDNA and disease association studies.

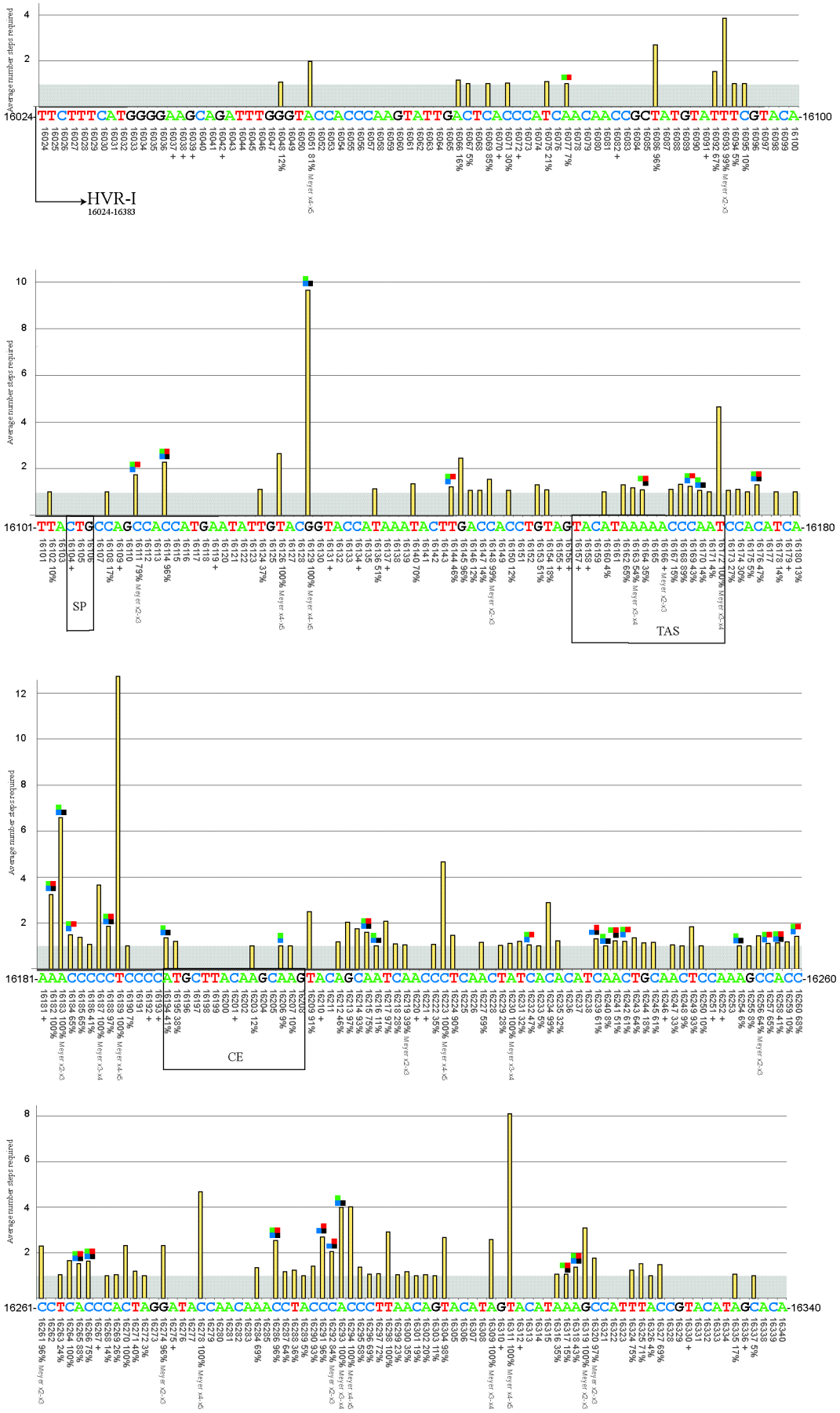
Mutation hotspots and conserved areas in the control region

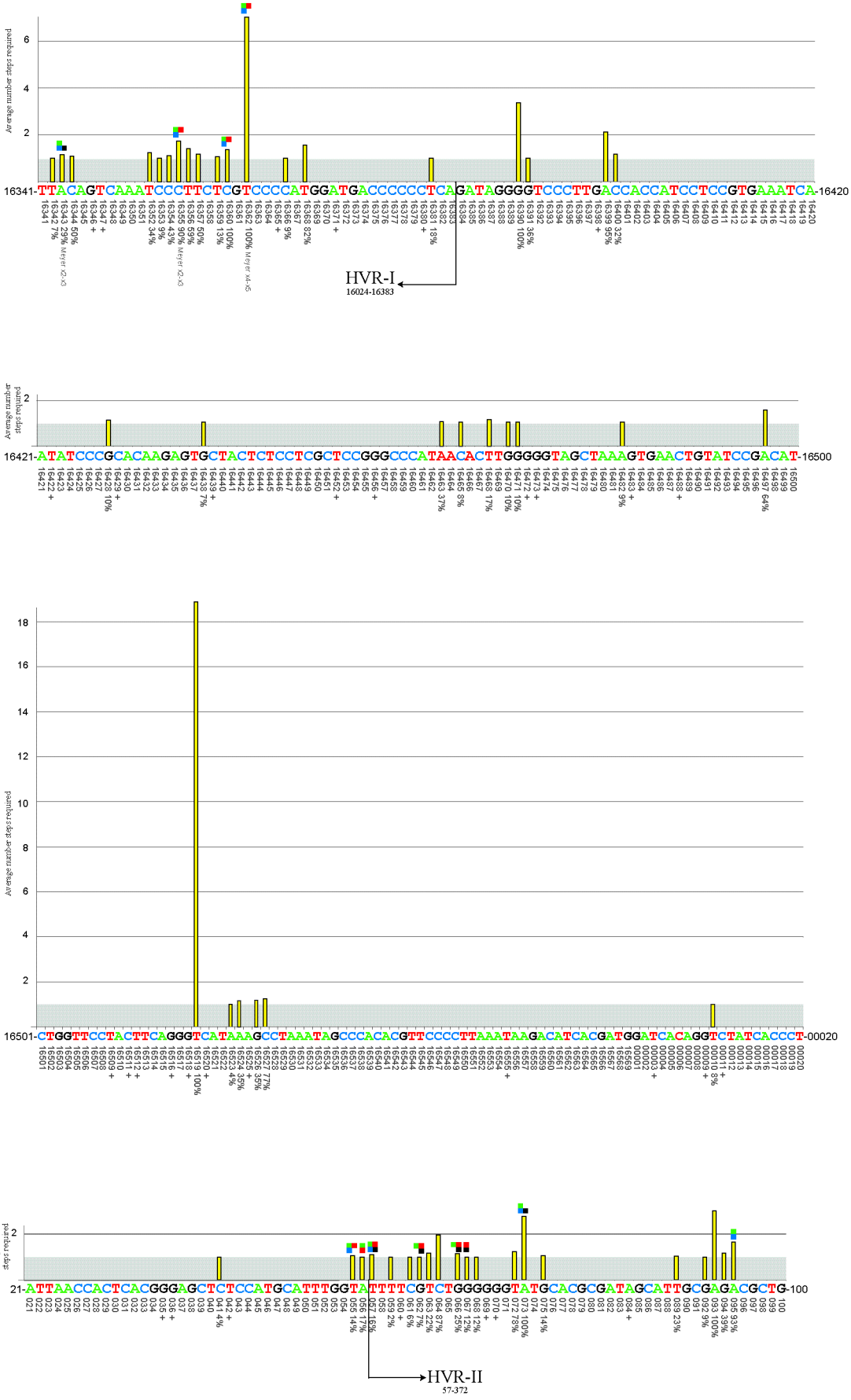
The average number of changes required to fit the control region bases to the coding region trees is shown in Figure 5.2. Each column represents the average changes for that character within the data sets in which it was informative, and the percentage of the total of 114 minimal data sets it was informative in is shown below the character name (Appendix E5.1 tabulates the average steps required over all data sets).

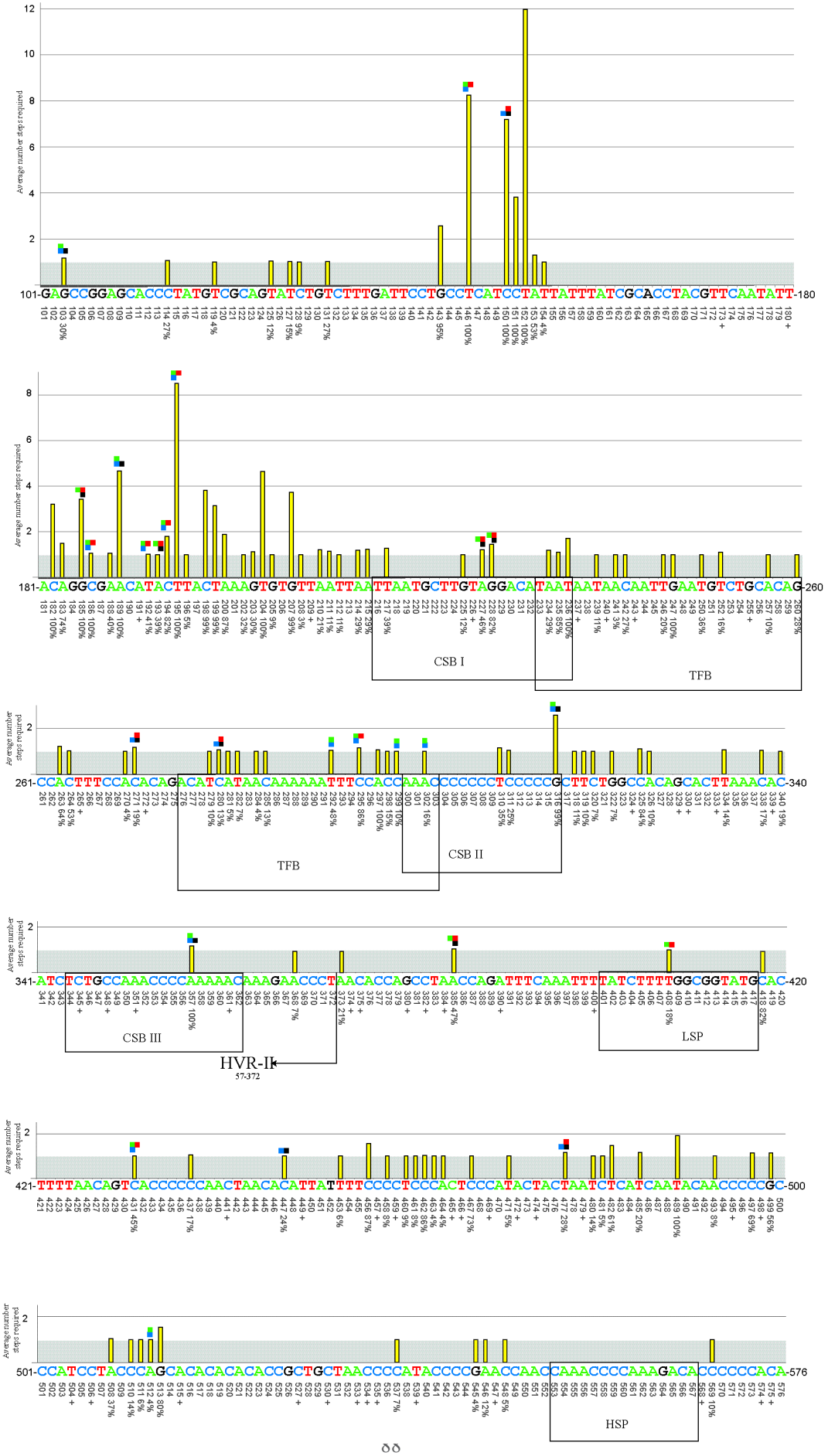
Figure 5.2 shows a number of bases have high rates of recurrent mutation compared to others in the control region. Ten bases require an average of more than six mutations to fit the 75-taxa coding region trees: nt16129, nt16183, nt16189, nt16311, nt16362, nt16519, nt146, nt150, nt152 and nt195. Transitions between cytosine and thymine at nt16519 show the highest rate of change, with more than 18 mutations required on average to fit it to the coding trees. The mutability at this base is in stark contrast to the conserved nature of the surrounding sequence, which is thought to include the origin of H-strand replication sites (Fish *et al.*

Figure 5.2 Chart of control region 'hotspots' (following pages)

The average number of steps required for each character in the control region when mapped to the 75-taxa datasets is shown (average is over data sets characters were informative in, not complete total number of data sets. The percentage of total data sets (n=114) that the base was informative in is shown following the base number. A '+' symbol indicates that there is variation in the global dataset at the base but it is not parsimony informative. HVR-I positions found to have a relative rate of more than 2 by Meyer *et al.* (1999: p1108) are marked 'Meyer', with the relative rate below the base number. The 7S DNA forming the displacement loop (D-loop) ranges from nt16106 to approximately nt191 (Fish *et al.* 2004, Meyer *et al.* 1999). The following features are marked by boxes: SP: trinucleotide stop-point for the 3' ends of the 7S DNA strands forming the D-loop (Meyer *et al.* 1999), TAS: termination associated sequence (Gemmell *et al.* 1996); CE: possible control element, Meyer *et al.* 1999); CSB: conserved sequence block, (Gemmell *et al.* 1996); TFB: mt transcription-factor binding sites (Meyer *et al.* 1999); LSP: light strand promoter region (Taanman 1999); HSP: heavy strand promoter (Taanman 1999). A character fully compatible with the coding region trees has an average step value of 1, and this is marked by grey shading. Bases undergoing transversions are indicated by squares above the column representing the states present at that position: red=thymine, green=adenine, blue=cytosine, black=guanine.


HVR-I
16024-16383
SP
TAS
CE


HVR-I
16024-16383
HVR-II
57-372


CSB I
TFB
TFB
CSB II
CSB III
HVR-II
57-372
LSP
HSP

2004, Taanman 1999) as most bases between nt16400 and nt50 show no variation at all.

Meyer *et al.* (1999) examined the relative rates of substitutions in the HVR-I from available control region sequences, finding some positions had rates of up to six times greater than average. Their results are indicated in Figure 5.2 and in general fit well with the results from this study, although there are some discrepancies: for example nt16126 and nt16129 fall within the 4-5 times greater rate in Meyer *et al.*'s study, but this analysis shows a far higher rate of change at nt16129 than nt16126. Other differences include the rate at nt16166, which is assigned a rate 2-3 greater than average by Meyer *et al.* but is not parsimony informative in this study.

The low rates of change in some parts of the control region (Figure 5.2) suggest these have an important functional role in mtDNA replication and/or transcription. Several features identified by human and comparative mammalian studies (Gemmell *et al.* 1996, Meyer *et al.* 1999, Taanman 1999) are shown in Figure 5.3, and this analysis also suggests a possible role for the conserved bases between nt155 and nt180, which are bordered by clusters of highly mutable sites.

5.3 Methods used for the coding vs. HVR-I analysis

Preparation of data sets

Eighteen haplogroups with strong support for monophyly in the consensus coding region tree (Appendix D4.1) were selected from the 'globalmtDNAcompletehaps.nex' data set of haplotype sequences (Table 5.2). The number of haplotypes within the haplogroups ranged from 16 (M/Q) to 261 (M/D), with a total of 1364

Haplogroup	n	Haplogroup	n
L1bc	30	N/A	56
L2	47	N/R/B4a	45
L3(bd)f	19	N/R/B4bd	28
L3ei	17	N/R/HV	254
M/D	261	N/R/JT	121
M/G	63	N/R/P	17
M/M7	88	N/R/R9	55
M/M8	51	N/R/U	164
M/Q	16	N/W	32

Table 5.2 Haplogroups used in coding region vs. HVR-I phylogeny comparison

sequences in the complete data set ('globalhapcoded.nex, digital Appendix F6.1).

The data set was extended by eighteen characters named for each of the haplogroups using MacClade (the original length was 16693, with character codes this increased to 16711 characters). Sequences belonging to a haplogroup were coded as 'A' for that character, and 'T' for the remaining 17 characters. 5000 datasets of 15 taxa were constructed from the globalhapcoded.nex data set by randomly choosing three of the 18 haplogroups and five sequences from within each using a C++ code (D. Bryant, Appendix C5.3).

PAUP* parsimony analysis

Parsimony searches were conducted for each of the 5000 data sets using a) the mtDNA coding sequence, b) the HVR-I sequence (nt16024-nt16383), c) the HVR-I sequence with weight set 1 and d) the HVR-I sequence with weight set 2.

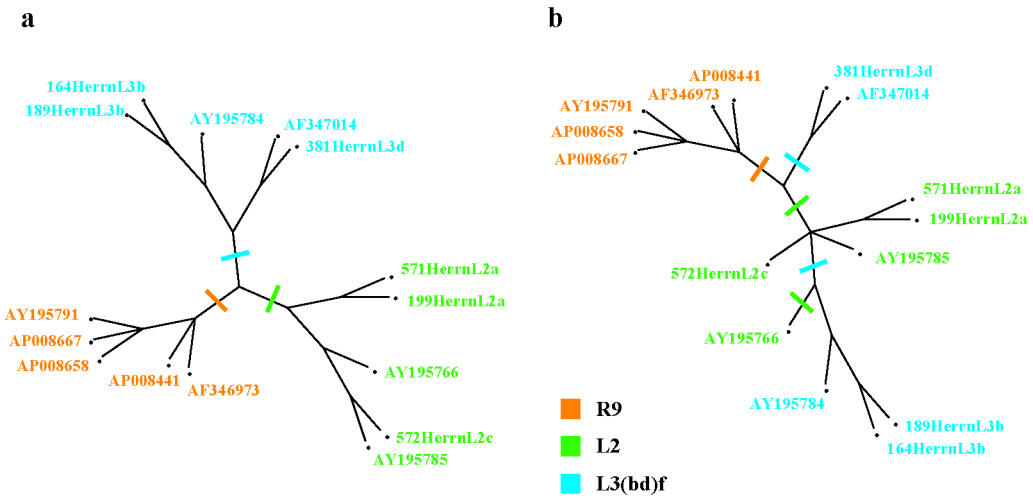


Figure 5.3 Example of parsimony haplogroup scoring for coding region vs HVR-I comparison
a) The single tree found from the coding region sequence of a random dataset of five sequences from haplogroups L2, L3(bd)f and N/R/R9 (76 informative characters, parsimony score=79). The coloured bars drawn across the tree branches indicate the fit to the tree of the haplogroup characters which were added to the sequence data. The score of this tree when the 18 haplogroup characters are mapped is the minimum of three; with the characters for L2, L3(bd)f and N/R/R9 each requiring a single step to fit the tree.
b) The single tree found for the same data set as in a), when the HVR-I sequence (nt16024-nt16383) is used to determine the phylogeny (number of informative characters=13, parsimony score=17). The fit of the haplogroup characters to this tree results in a score of 5, as 2 steps are required for haplogroups L2 and L3(bd)f. The maximum parsimony score for any haplogroup character in the data sets of three random haplogroup sets each containing five haplotypes is five, and the maximum score for a tree with the haplogroup characters mapped is 15.

The weight sets were determined from the results of the 75-taxon analysis, using the average number of steps required to fit the coding trees they were informative in, (Figure 5.1, Appendix E.5.1), as a measure of their relative mutation rates. The average number of steps was rounded to the nearest whole number for each character and in the first set the weights were calculated by dividing one by the number of steps, resulting in nine weight categories from 0.08 to 1. The second weight set was more conservative; when characters required an average of two or more steps to fit the coding trees in the 75-taxon analysis, 0.1 was subtracted for each additional step. This weight set had nine categories, from 0 to 1 (no characters had an average of six steps to fit a 0.5 category).

The PAUP* analysis was designed to output a single log file of details of the heuristic parsimony search results, and four text files with details of the lengths the haplogroup characters required on the first tree found for each of the four tree searches conducted on each data set. The PAUP* commands are included in Appendix C5.4.

The parsimony score files for the haplogroup characters on the first tree found for each of the 5000 replicates were edited using BBedit Lite and imported into an Excel® workbook. An example of how the haplogroup character parsimony scores reflect the differences in tree reconstructions from the two parts of the mtDNA molecule (coding and HVR-I) is given in Figure 5.3.

5.4 Results of the coding vs. HVR-I analysis

Figure 5.4 charts the results of the coding and HVR-I comparative phylogeny analysis. The unweighted control region characters predicted the coding region tree precisely (haplogroup character score=3) in 1020 of the 5000 replicates, with both weight sets showing a slight improvement in this category (in weight set 1 1094 trees had haplogroup score 3, and in weight set 2 1090 trees scored 3). Almost half of the HVR-I data sets had haplogroup scores of 4, with a gradual decline in numbers of data sets with higher haplogroup character scores to the maximum of 13, which occurred in a single replicate.

The weight sets did not have a strong effect on the resolution to haplogroup of the HVR-I data sets; while a slight improvement was seen in attaining a haplogroup score of 3, there was no marked improvement in performance seen in the remainder of the categories.

Table 5.3 details the results of the analysis when each of the 18 haplogroups is considered separately.

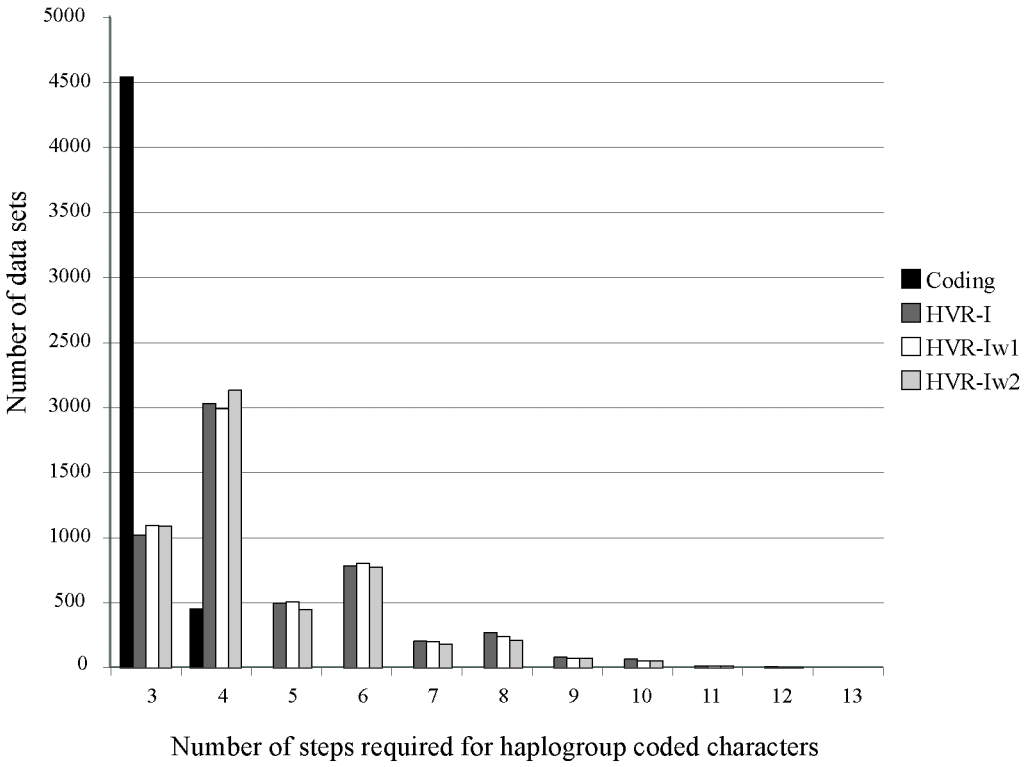


Figure 5.4 Haplogroup steps required for trees coding and HVR-I
HVR-1w1: The HVR-I region, nt16024-nt16383, with weight set 1
HVR-1w2: The HVR-I region with weight set 2.

The minimum parsimony score for each haplogroup on a tree is 1, and the maximum 5 (Figure 5.2). The proportions of steps required for each haplogroup character on the coding and unweighted HVR-I trees highlights the differences between the haplogroups in terms of reconstructing the coding region phylogeny from the HVR-I sequence. The haplogroups which required two steps or more in over 50% of the HVR-I 5000 replicates are shaded.

The combination of haplogroups which produced the highest haplogroup parsimony score was N/R/P, N/R/U and N/R/HV, which occurred four times in the 5000 replicates (the probability of selecting any combination of three haplogroups is 1 in 4896), with scores of 7, 10, 12 and 13. Other haplogroup combinations requiring more than 10 steps were N/R/HV, N/R/U and N/R/R9, which occurred five times, with scores of 6, 7, 10, 11 and 12. N/R/P Samples from the Trobriand Islands sequenced for this study (Chapter Two) were found to have an identical sequence to the N/R/HV rCRS sequence, and the low resolution of these two haplogroups from the HVR-I sequences seen in this analysis reflects the lack of HVR-I changes seen in the higher branches of these phylogenies leading from the ancestral macrohaplogroup N vertex to the HV and P

Table 5.3 Coding region vs. HVR-I phylogeny comparison: haplogroup results
Haplogroups requiring two or more steps in >50% of the HVR-I replicates are shaded.

	Steps	Coding	HVR-I
L1bc	1	1.00	0.95
	2	0.00	0.05
	3	0.00	0.00
	4	0.00	0.00
	5	0.00	0.00
L3ei	1	1.00	0.12
	2	0.00	0.52
	3	0.00	0.29
	4	0.00	0.07
	5	0.00	0.00
M/M8	1	1.00	0.76
	2	0.00	0.22
	3	0.00	0.02
	4	0.00	0.00
	5	0.00	0.00
N/A	1	1.00	0.90
	2	0.00	0.10
	3	0.00	0.00
	4	0.00	0.00
	5	0.00	0.00
N/R/HV	1	1.00	0.42
	2	0.00	0.33
	3	0.00	0.12
	4	0.00	0.10
	5	0.00	0.02
N/R/R9	1	1.00	0.56
	2	0.00	0.39
	3	0.00	0.04
	4	0.00	0.00
	5	0.00	0.00
L2	1	1.00	0.55
	2	0.00	0.40
	3	0.00	0.06
	4	0.00	0.01
	5	0.00	0.00
M/G	1	1.00	0.46
	2	0.00	0.38
	3	0.00	0.13
	4	0.00	0.02
	5	0.00	0.00
M/M7	1	1.00	0.18
	2	0.00	0.61
	3	0.00	0.19
	4	0.00	0.02
	5	0.00	0.00
N/R/B4a	1	1.00	0.98
	2	0.00	0.02
	3	0.00	0.00
	4	0.00	0.00
	5	0.00	0.00
N/R/JT	1	1.00	0.88
	2	0.00	0.12
	3	0.00	0.00
	4	0.00	0.00
	5	0.00	0.00
N/R/U	1	1.00	0.24
	2	0.00	0.47
	3	0.00	0.20
	4	0.00	0.08
	5	0.00	0.01
L3(bd)f	1	0.50	0.18
	2	0.50	0.50
	3	0.00	0.28
	4	0.00	0.03
	5	0.00	0.00
M/D	1	1.00	0.41
	2	0.00	0.36
	3	0.00	0.15
	4	0.00	0.06
	5	0.00	0.02
M/Q	1	1.00	0.91
	2	0.00	0.09
	3	0.00	0.00
	4	0.00	0.00
	5	0.00	0.00
N/R/B4bd	1	0.99	0.85
	2	0.01	0.15
	3	0.00	0.00
	4	0.00	0.00
	5	0.00	0.00
N/R/P	1	0.98	0.31
	2	0.02	0.43
	3	0.00	0.17
	4	0.00	0.07
	5	0.00	0.01
N/W	1	1.00	0.74
	2	0.00	0.23
	3	0.00	0.03
	4	0.00	0.00
	5	0.00	0.00

descendants.

As the N/R/P haplogroup is geographically restricted at present to Near Oceania the likelihood of misassignment of samples to this haplogroup using HVR-I sequences is relatively low in this case. However it is interesting to note that the N/R/HV and N/R/U haplogroups also appear difficult to differentiate using HVR-I sequences as both of these groups are common in European populations, and outside of Europe in populations of European descent.

5.5 Discussion

The results of the analysis of homoplasy, in both the coding and control regions, are very interesting, inspiring many further questions regarding the generation of the rate diversity seen in the control region, and the potential causes of high mutability at particular bases in the coding region. The 75-taxa analysis undertaken here provides a relative scale for the likelihood of repeated polymorphisms at a single site, on a global scale, but may be adversely influenced by the predominance of Eurasian haplotypes in the existing global data set, particularly in N/R/HV and M/D haplogroups (Table 4.3). An improvement to this analysis would be to refine the selection of haplotypes to make up the random 75 taxa data sets, taking into account their distribution within the phylogeny in a less arbitrary way than the division to L3 and non-L3 groups used here.

Arguments for hypervariability over recombination in explaining the presence of homoplasy in any particular data set are intuitively strengthened if these positions are known to be homoplastic in a global as well as a local sense. An application of the information gained from this analysis is the assessment of the homoplasmic sites seen in the phylogenies generated in Chapter Three. The N/R/P haplogroup in particular (Figure 3.2) seemed to have an improbably high number of homoplasmic characters when the P sequences were analysed with the single R12 and R21 sequences. Four of the sequences in the analysis of the N/R/P haplogroup (DQ112752, AY963584, AY289054, DQ404446, far left, Figure 3.2d) caused conflict between trees in Oceanic phylogenetic analyses (Figure 2.4) due to the combination of states they have at five coding region bases: nt10398, nt11404, nt12361, nt15613 and nt15607.

The present analysis has demonstrated that homoplasy at nt10398 is common, with an average of 4.2 mutations at this base required to fit the random 75-taxa phylogenies. However, the nt11404, nt12361, nt15613 bases had the base value of an average of one step to fit the trees in the 75-taxa analysis, (nt15607 required 1.5 on average, Appendix E5.1). Two of these four sequences (R12 and R21) represent single examples of haplogroups, while two have the nt15607G transition and are thus assigned to haplogroup N/R/P, but share no other variants with P sequences. It is possible that the tangled patterns of changes seen within these sequences are due to human errors, many of which have been documented from previous studies (for example Bandelt *et al.* 2002, Bandelt *et al.* 2004, Bandelt *et al.* 2007). An alternative explanation exists however, which does not require the coincidence of multiple relatively rare mutations seen in the phylogenetic reconstruction. If heteroplasmy persists over a number of generations, and intra-lineage recombination occurs, a pattern could emerge like the one seen in the R12, R21 and P lineages. This

hypothesis predicts that additional sequences from these haplogroups would place the homoplastic polymorphisms high in the branches, as the inclusion of the nt15607 variant in the set dates the postulated occurrence of intra-lineage recombination to before the split to the many P lineages which occurs in the phylogeny directly after the macrohaplogroup R vertex. Testing this theory with more samples from these haplogroups is an aim for future research.

In the previous chapter several data sets were identified as having significantly higher levels of homoplasy than expected when tested using the phi statistic (Bruen *et al.* 2006). Homoplastic bases were identified in one of the random data sets of 30 individuals generated from the global data set (Random_30, Figure 4.8). Twenty-six of the 56 characters which required an average of more than one step to fit the most parsimonious trees came from the control region; the thirty coding region bases are listed in Table 5.4 with the number of steps they required for both the Random_30 and 75-taxa analyses.

Table 5.4 Homoplastic bases in Random_30 data set and 75-taxa steps

Character number	Steps in Random_30	Steps in 75-taxa
629	2	<1.5
709	2	6.6
1442	3	<1.5
1811	2	<1.5
2706	2	<1.5
3010	3.6	3.88
3397	2	<1.5
3666	2	<1.5
4216	2	<1.5
4386	2	<1.5
5417	2	<1.5
5460	2	3.75
7389	2	<1.5
8701	2	1.59
10398	2	4.18
10685	2	<1.5
10810	2	<1.5
11467	2	<1.5
11719	2	<1.5
11969	2	<1.5
12007	2	2.23
12308	2	<1.5
12372	2	1.65
12705	2.2	1.75
13105	4	2.13
13708	3	2.87
14178	2	<1.5
14798	2	1.82
15067	2	<1.5
15607	2	1.51

Twelve of the homoplastic bases identified in the Random_30 data set had scores of more than 1.5 in the 75 taxa analysis., including all but one of the changes seen with more than two steps in the Random_30 data set. Character nt1442 requires three steps in the Random_30 data set, and 1.4 in the 75-taxa analysis in which it appeared in 99% of the data sets generated (Appendix E5.1). So, in this case 'hypervariability' may account for almost half of the homoplasy observed in the coding region.

The visual disparity in rates of change found in the control region from this study, particularly in hypervariable regions I and II (Figure 5.2), caution against the use of this part of the mtDNA sequence for phylogenetic reconstruction and dating estimates without consideration of the variation observed. While these kinds of study were standard in the 1990s, it is common nowadays to apply a more conservative approach to HVR-I sequence information through additional typing of coding region polymorphisms (for

an Oceanic example of this approach see Friedlaender *et al.* 2007). The continued use of HVR-I sequence data to explore issues in prehistory is naive given the existing knowledge of rate variation in the control region, particularly when used to date ancestral vertices (Kayser *et al.* 2006, Hill *et al.* 2007).

The comparison of trees obtained from coding and HVR-I regions provides a relative scale of 'strength' of each haplogroup in terms of the reconstruction of trees from HVR-I characters, but disappointingly did not show any clear gains from applying the two weight sets devised from the results of the 75-taxa analysis. Refining the weighting system in future may improve these results. In the analysis of the large data sets of HVR-I sequences which follows in Chapter Six a conservative approach is taken, working back from the entire mt genome phylogenies to identify haplotypes and construct tentative HVR-I phylogenies for major haplogroups present in Oceania.

6. Collection and analysis of new Polynesian HVR-I samples in an Oceanic context

This chapter describes the HVR-I (hypervariable region 1) sequencing and analysis of 46 Polynesian samples within a large data set of more than 4000 sequences collected from public databases. The whole genome phylogenies reconstructed in Chapter Three were used to identify coding region polymorphisms to target in further analyses of the sample set, and enabled precise branching points from the N/R/B4a phylogeny to be determined without needing to sequence the entire mtDNA molecule. In Chapter Five it was shown that the success rate of recovery of trees from HVR-I data showing the same topology as more robust coding region trees, varied widely between haplogroups. As the weighting schemes devised for the HVR-I did not show any significant improvement over the unweighted case a conservative approach was taken in this chapter to the analysis of the new sample set within the larger Asian and Oceanic context. Haplotypes shared within and between groups were identified, and haplogroups assigned according to information contained in the whole mt genome phylogenies.

6.1 Collection of HVR-I and coding SNP data from 46 Polynesian samples

Sample collection

A set of 47 finger prick blood and/or buccal swab samples from volunteers of Polynesian descent (and one New Zealand European control) living in Auckland was collected in 2005 by Brad Fris, for a Y-chromosome variation analysis carried out as a Master of Science degree project at the University of Auckland in conjunction with the Institute of Environmental Science and Research (Fris 2006). All volunteers gave their informed written consent for the study which was approved by the University of Auckland Human Subjects Ethics Committee (reference number 2005/082). DNA was extracted using a Chelex method, and the samples were typed for a number of Y-chromosome polymorphisms (Fris 2006) to aid in the development of a Y STR/SNP system for forensic casework in New Zealand. To complement the Y chromosome analysis I was approached to undertake mtDNA control region sequencing.

mtDNA sequencing

As initial attempts to amplify the long mtDNA fragment 2 (Appendix C2.1) from the samples were not successful, a 600bp product encompassing the mtDNA sequence from nt15900 to nt16500 was amplified using 23F and 23R primers (Rieder *et al.* 1998, Appendix B and C2.1). The PCR conditions used for this amplification were the same as those used for the segment amplifications in the whole mt genome methodology, described in Appendix C2.1.

20F and 20R primers were used to amplify a product including the nt14022 base which defines the B4a1a1 sub-group in samples with the immediate pre-Polynesian motif HVR-I sequence (N/R/B4a). A subset of the samples (n=13) were also typed for the nt6905A/G polymorphism using primers 10F and 10R. Sequencing methodology used was as described in Appendix C.2, with the exception of the typing for the nt6905 polymorphism, for which both the sequencing reactions and separation of fragments was carried out by the Allan Wilson Genome Service, Albany Campus, using a ABI3730 Genetic Analyzer.

Results and discussion

Sequence data from nt16024-nt16569 of the control region, including the HVR-I, was obtained from the 46 Polynesian samples (Table 6.1). The new sequences include 11 samples with Maori ancestry, and these increase the number of haplotypes reported from New Zealand from 8 to 10 (Sykes *et al.* 1995, Murray-McIntosh *et al.* 1998, Whyte *et al.* 2005). Sixteen of the 17 haplotypes belonged to the common Polynesian mtDNA haplogroup N/R/B4a (Figure 3.4). The most common haplotype (23 individuals) was the 'Polynesian motif' (corresponding to entire mt haplogroup N/R/B4a1a1PM) without variants. Three samples had this motif with an additional transition at nt16092, and two had a transition at nt16051. Eleven haplotypes found in single individuals also had the B4a1a1PM haplotype with additional polymorphisms. The single sample which did not belong to the N/R/B4a haplogroup (EF077389) came from an individual with ancestry in the Solomon Islands, and appears to belong to haplogroup M/M27b (due to shared polymorphisms at nt16209C, nt16299G, nt16390A and nt16519C) which has not been reported from Remote Oceania but is relatively common in parts of Near Oceania (Merriwether *et al.* 2005, Friedlaender *et al.* 2007).

The remaining six individuals had the immediate pre-Polynesian Motif HVR-I signature (which corresponds to the whole mtDNA vertex N/R/B4a); five with no additional polymorphisms and one with a transition at nt16311, and came from individuals with Fijian, Maori, Tongan and Samoan ancestry. The phylogeny determined in Chapter Three (Figure 3.4) revealed the distance between the transition at nt16247 which separates the HVR-I pre-Polynesian Motif (16217C, 16261T) from the full Polynesian Motif (16217C, 16247G, 16261T) is quite large, with several coding region polymorphisms occurring along the branches between the two events. To assess where the six Polynesian samples with the HVR-I pre-Polynesian Motif branch from the B4a tree they were tested for the polymorphism at nt14022, which defines the B4a1a1 subhaplogroup, with subsequent hierarchical testing of B4a1a (at nts 6719, 12239 and 15746) and B4a1 (nt10238) polymorphisms planned if necessary. All six of the samples (EFO77363, EFO77368-EFO77369, EFO77374, EFO77391-EFO77392) had the transition to guanine at nt14022 which places them clearly within the B4a1a1 part of the B4a phylogeny.

Table 6.1 Auckland sample set details and results

Accession	Sample name	Ethnicity	Haplotype	Differences to rCRS 16024-16569	Other sites
EFO77357	1	Samoan	B4a1a1PM	16092C, 16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16519C	*6905A*
EFO77358	2	Samoan	B4a1a1PM	16182C, 16183C, 16189C, 16217C, **16247A/G**, 16261T, 16519C	*6905A*
EFO77359	3	Maori	B4a1a1PM	16051G, 16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16519C	**6905G**
EFO77360	4	Maori	B4a1a1PM	16092C, 16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16519C	*6905A*
EFO77361	5	Maori	B4a1a1PM	16086C, 16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16519C	*6905A*
EFO77362	6	Cook Is	B4a1a1PM	16051G, 16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16519C	**6905G**
EFO77363	7	Fijian	B4a	16182C, 16183C, 16189C, 16217C, 16261T, 16311C, 16519C	**6905G, 14022G**
EFO77364	8	Samoan	B4a1a1PM	16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16263C, 16519C	*6905A*
EFO77365	9	Maori	B4a1a1PM	16092C, 16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16519C	*6905A*
-	10	NZ European	n.d.	n.d.	
EFO77366	11	Samoan	B4a1a1PM	16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16519C	*6905A*
EFO77367	12	Maori	B4a1a1PM	16182C, 16183C, 16189C, 16217C, **16247A/G**, 16261T, 16519C	*6905A*
EFO77368	13	Maori	B4a	16182C, 16183C, 16189C, 16217C, 16261T, 16519C	*6905A*, **14022G**
EFO77369	14	Maori	B4a	16182C, 16183C, 16189C, 16217C, 16261T, 16519C	**6905G, 14022G**
EFO77370	15	Tongan	B4a1a1PM	16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16311C, 16519C	
EFO77371	16	Tokelauan	B4a1a1PM	16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16519C	
EFO77372	17	Samoan	B4a1a1PM	16182C, 16183C, 16189C, 16217C, 16247G, 16249C, 16261T, 16519C	
EFO77373	18	Samoan	B4a1a1PM	16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16519C	
EFO77374	19	Tongan	B4a	16182C, 16183C, 16189C, 16217C, 16261T, 16519C	**14022G**
EFO77375	20	Tongan	B4a1a1PM	16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16519C	
EFO77376	21	Cook Is	B4a1a1PM	16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16519C	
EFO77377	22	Cook Is	B4a1a1PM	16182C, 16183C, 16189C, 16217C, 16247G, 16261T	
EFO77378	23	Maori	B4a1a1PM	16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16519C	
EFO77379	24	Samoan	B4a1a1PM	16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16519C	
EFO77380	25	Samoan	B4a1a1PM	16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16519C	
EFO77381	26	Cook Is	B4a1a1PM	16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16291T, 16519C	
EFO77382	27	Samoan	B4a1a1PM	16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16519C	
EFO77383	28	Samoan	B4a1a1PM	16181C, 16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16519C	
EFO77384	29	Tongan	B4a1a1PM	16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16519C	
EFO77385	30	Maori	B4a1a1PM	16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16519C	
EFO77386	31	Niuean	B4a1a1PM	16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16519C	
EFO77387	32	Cook Is	B4a1a1PM	16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16519C	
EFO77388	33	Cook Is	B4a1a1PM	16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16519C	
EFO77389	34	Solomon Is	M/M27b[1]	16086C, 16209C, 16223T, 16299G, 16390A, 16519C	
EFO77390	35	Samoan	B4a1a1PM	16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16519C	
EFO77391	36	Samoan	B4a	16182C, 16183C, 16189C, 16217C, 16261T, 16519C	**14022G**
EFO77392	37	Tongan	B4a	16182C, 16183C, 16189C, 16217C, 16261T, 16519C	**14022G**
EFO77393	38	Maori	B4a1a1PM	16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16519C	
EFO77394	39	Cook Is	B4a1a1PM	16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16519C	
EFO77395	40	Samoan	B4a1a1PM	16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16465T, 16519C	
EFO77396	41	Samoan	B4a1a1PM	16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16519C	
EFO77397	42	Tongan	B4a1a1PM	16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16519C	
EFO77398	43	Tongan	B4a1a1PM	16163G, 16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16519C	
EFO77399	44	Maori	B4a1a1PM	16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16519C	
EFO77400	45	Samoan	B4a1a1PM	16093C, 16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16519C	
EFO77401	46	Tongan	B4a1a1PM	16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16519C	
EFO77402	47	Cook Is	B4a1a1PM	16182C, 16183C, 16189C, 16217C, 16247G, 16261T, 16519C	

The sequence data extended into the tRNA coding region flanking the 5' end of the control region, revealing that none of the samples had the transition at nt15924 which is seen in the B4a1a1 whole mt genome sequences DQ372877, DQ372874 and DQ372875 from the Marshall Islands and Kapingamarangi.

A subset of the samples were examined for the presence of the nt6905G transition that occurs in three of the whole mt genomes sequences (AY289069, AY289083, AY289093; from the Cook Islands, Papua New Guinea and Samoa respectively (Figure 3.4). Four of the 13 samples tested had nt6905G (EFO77359, EFO77362-EFO77363, EFO77369; from individuals of Maori, Cook Island and Fijian descent), and it is interesting to note that two of these have the full Polynesian Motif HVR-I sequence, while the remaining two have the immediate pre-Polynesian Motif. This pattern was also seen in the whole mt genomes, where in the labelled phylogeny (Figure 3.4) a reversion of the nt16247G transition has been inferred in sequence AY289083.

It was observed from electropherograms that two of the Auckland samples (EFO77358 and EFO77367, individuals with Samoan and Maori ancestry) showed evidence of heteroplasmy at nt16247. As the nt16247G variant occurs in populations throughout Oceania, and at high frequencies particularly in Polynesia, it is likely to have arisen before the colonisation of Remote Oceania. The apparent heteroplasmy seen in the Auckland samples may be the result of incomplete fixation of the original nt16247A mtDNA molecule and the 'new' molecule carrying the nt16247G variant; or evidence of more recent recurrent mutations at nt16247 following the first transition event to guanine. The most parsimonious explanation, that the heteroplasmy generated by the initial transition event has persisted to the present requires the inheritance of a mixed population of mtDNA molecules over a long timeframe. A conservative estimate of 1200 years for the common ancestry of individuals with the Polynesian motif haplotype would require the persistence of mtDNA heteroplasmy through 40 generations at 30 years per generation, (60 generations at 20 years per generation).

6.2 Compilation of HVR-I data sets from Oceania, Asia and the Americas

Sequences covering the first hypervariable region (HVR-I, nt16065-nt16373) from Oceania, Asia and the Americas were identified by literature and NCBI database searching and aligned using SE-AL (version 2.0a11, Rambaut 1996). Different treatments of the length extensions and deletions in the cytosine repeat region between nt16184 and nt16193 made alignment of this region difficult, requiring two bases, nt16192 and nt16193, to be excluded. The HVR-I portion of the sequence was also extracted from geographically relevant mt genomes from the global data set described in Chapter Four, with a subset of 192 sequences

Table 6.2 HVR-I nt16065-16373 data set accession details
(excl: excluded, ambig: ambiguities present, *: entire mt genomes)

Accession numbers from	to	n	Notes	Authors
AB048147	AB048181	20	15 excl. (ambig.)	Oota et al 2002
AB093862	AB094026	165		Tajima et al 2004
AB119285	AB119445	161		Ohashi et al 2006
AB122153	AB122706	554		Tajima et al 2004
AF176125	AF176199	27	4 excl. (ambig.)	Redd and Stoneking 1999
AF346970	AF347013	17*		Ingman et al 2000
AF382012		1*		Maca-Meyer et al 2001
AF392063	AF392434	364	8 excl. (ambig.)	Yao et al 2002
AJ634960	AJ635160	173	26 excl. (gaps)	Tommaseo-Ponzetta et al 2002
AJ842744	AJ842751	8*		Trejaut et al 2005
AJ967036	AJ967675	640	1 excl. (ambig.)	Trejaut et al 2005
AP008537	AP008728	192*		Tanaka et al 2004
AY195748	AY195792	17*		Mishmar et al 2003
AY255133	AY255180	48*		Kong et al 2003
AY289068	AY289102	31*		Ingman and Gyllensten 2003
AY519484	AY519497	14*		Starikovskaya et al 2005
AY546723	AY547259	536	40 excl.	Wen et al 2005
AY570524	AY570526	3*		Starikovskaya et al 2005
AY604071	AY604153	18	2 excl. (ambig.)	Whyte et al 2005
AY615359	AY615361	3*		Starikovskaya et al 2005
AY950286	AY950300	15*		Thangaraj et al 2005
AY956412	AY956414	3*		Friedlaender et al 2005
AY963572	AY963584	13*		Macaulay et al 2005
D84723	D84773	51		Horai et al 1996
DQ137398	DQ137411	14*		Merriwether et al 2005
DQ144507	DQ144539	30	3 excl. (ambig.)	Lewis et al 2005
DQ145827	DQ149065	141		Luangtrakool,K., Sanpachudayan,T., Tharaphan,P., Suphavilai,R., Srisawat,C., Suktitipat,B., Poolsuwan,S. and Lertrit,P., Direct submission (27-JUL-2005)
DQ303483	DQ303578	96		Helgason et al 2006
DQ309847	DQ309867	21		Ricaut,F.-X., Arganini,C., Staugton,J., Bellatti,M., Lahr,M.-M.
DQ372868	DQ372886	19		Pierson et al 2006
DQ418488		1*		Li,S. Direct Submission (27-FEB-2006)
DQ437577		1*		Li,S. Direct Submission (07-MAR-2006)
DQ462232	DQ462234	3*		Li,S. Direct Submission (27-MAR-2006)
EF068416	EF069259	704	62 excl. (ambig.) & repeats	Hill et al 2007
EF077357	EF077402	45		Pierson,M.J., Gemmell, N. and Fris,B. Direct Submission (23-OCT-2006)
U25374	U25413	19	41 excl. (ambig.)	Redd et al 1995
U47164	U47271	153	Several excl. (gaps)	Sykes et al 1995
	Total	4321		

taken from the large Japanese Tanaka *et al.* (2004) data set. Table 6.2 lists details of the total 4321 sequences collected.

Each sequence was given a two to six letter suffix according to geographic details provided in the GenBank file, and assigned to one of 13 geographic regions for analysis using Arlequin (version 3.01, Excoffier *et al.* 2005). Sequences from New Zealand residents who have stated ancestry outside of New Zealand were labelled with 'NZ' before the geographic code for the place of ancestry, for example 'NZCook' for a New Zealander of Cook Islands descent, and in one case assigned to a region other than Remote Oceania (when ancestry was in the Solomon Islands). The sequences from Sykes *et al.* (1995) were deposited as haplotypes in GenBank, and these have been expanded according to details given in their paper (Sykes *et al.* 1995: Table 2). The sequence name was made unique to the data set by adding a number to the geographical suffix following the accession number. For example 'haplotype 1' from Sykes *et al.* (1995), corresponding to accession U47145, is reported from three individuals and these are labelled U47145Taiw1, U47145Taiw2 and U47145Phil1 here. The composition of the data set by geographic region is summarised in Table 6.3.

The number of Remote Oceanic samples in the nt16065-nt16373 data set (n=211) was limited by sequence length, as large numbers of shorter sequences from Remote Oceania are available from analyses by Lum *et al.* (1998), Lum and Cann (2000) and Sykes *et al.* (1995) beginning from ~nt16189. Earlier attempts to use Arlequin to obtain haplotype information from the data set when sequences contained missing or ambiguous bases had inconsistent results, and consequently these sequences were removed. In order to analyse these shorter sequences they were added to an Oceanic subset of the nt16065-nt16373 data set, (excluding nt16192 and nt16193 as above), resulting in a second data set containing sites between nt16189-nt16370,and a total of 1191 Oceanic sequences. In this data set the sequences were divided to 19 regions: Near Guinea and Other Near Oceania as in the longer data set with the additional 602 sequences from Remote Oceania assigned to island group origin where possible (Table 6.4). The Arlequin format files for both the long (nt16065-nt16373) and short (nt16189-nt16370) HVR-I data sets are included within electronic Appendix F6.1 which also contains an Excel® workbook with the formatted results of the Arlequin analyses.

6.3 The distribution of HVR-I haplotypes in Oceania

Arlequin (version 3.01, Excoffier *et al.* 2005) was used to infer haplotypes from a distance matrix for each of the data sets, with each haplotype named for one of the sequences it contained. Summary details of the number of haplotypes within each region are listed in Table 6.5, for the nt16065-nt16373 Oceania, Asia and Americas HVR-I data set, and in Table 6.6 for the nt16189-nt16370 Oceanic HVR-I data set.

Table 6.3 HVR-I nt16065-16373 data set geographic details

n	Suffix	Source	Region
5	Amer	Americas	Americas n=134
96	Cana	Canada	
30	Peru	Peru	
1	Pima	Piman speaker	
2	Wara	Warao speaker	
84	Banj	Banjamasin	Borneo n=170
60	Kota	Kota Kinabalu	
26	Saba	Sabah	
31	Bai	Bai, China	East Asia n=1285
63	Buri	Buriat, Siberia	
7	Chin	China	
1	Chuk	Chukchi, Siberia	
38	Dai	Dai, China	
1	Daur	Daur, China	
19	EAsi	East Asia	
37	Ewen	Ewenki, Siberia	
43	Han	Han, China	
166	Hmon	Hmong, China	
1	Khir	Khirgiz	
1	Kore	Korea	
110	Kory	Koryak, Siberia	
37	Lisu	Lisu, China	
2	Miao	Miao, China	
363	Mien	Mien, China	
17	Mong	Mongolia	
60	NHan	Northern Han	
57	Nivk	Nivkhi, Sakhalin	
30	Nu	Nu, China	
29	Sali	Sali, China	
19	Sibe	Siberia	
40	Tibe	Tibet	
30	Tu	Tu, China	
83	Zhua	Zhuang, China	
35	Alor	Alor	Indonesia n=584
41	Ambo	Ambon	
89	Bali	Bali	
83	Indo	Indonesia	
86	Mana	Manado, Sulawesi	
40	Mata	Mataram, Lombok	
30	Palu	Palu, Sulawesi	
30	Teng	Tengger, Java	
60	Tora	Toraja, Sulawesi	
44	Ujun	Ujnug Padang	
46	Wain	Waingapu, Sumba	
51	Ainu	Ainu, Japan	Japan n=476
82	Hons	Honshu	
194	Japa	Japan	
104	Kyus	Kyushu	
45	Okin	Okinawa	

n	Suffix	Source	Region
53	Mala	Malaysia	Malaysia n=61
8	Oran	Orang Asli	
59	Gidr	Gidra, sw PNG	New Guinea n=299
173	Iria	Irian Jaya	
21	Kark	Karkar Island	
42	PNG	Papua New Guinea	
4	Trob	Trobriand Islands	
10	Anda	Andaman Islands	Other Island SEA n=15
5	Nico	Nicobar Islands	
59	Balo	Balopa Is., Manus	Other Near Oceania n=79
14	Bism	Bismarck Arch.	
5	Bour	Bourgainville	
1	NZSolo	Solomon Is., NZ	
146	Phil	Philippine Islands	Philippine Islands
3	Aust	Australes	Remote Oceania n=211
23	Cook	Cook Islands	
20	Haap	Ha'apai Is., Tonga	
2	Kapa	Kapingamarangi	
6	Marq	Marquesas	
6	Mars	Marshall Islands	
8	NZCook	Cook Islands, NZ	
1	NZFiji	Fiji, NZ	
25	NZMaor	Maori	
2	NZNiue	Niue, NZ	
1	NZPoly	Polynesian, NZ	
16	NZSamo	Samoa, NZ	
1	NZToke	Tokelau, NZ	
10	NZTong	Tonga, NZ	
6	Samo	Samoa	
4	Tahi	Tahiti	
30	Tong	Tonga	
47	Vanu	Vanuatu	
1	Camb	Cambodia	Southeast Asia n=197
176	Thai	Thailand	
20	Viet	Vietnam	
98	Amis	Amis	Taiwan n=664
109	Atal	Atayal	
89	Bunu	Bunun	
55	Paiw	Paiwan	
52	Puyu	Puyuma	
50	Ruka	Rukai	
63	Sais	Saisiat	
24	Taiw	Taiwan	
60	Tsou	Tsou	
64	Yami	Yami	

Table 6.4 HVR-I nt16189-16370 data set geographic details

Regions nt16189-nt16370	n	Additional suffix codes	Regions nt16065-nt16373	n
New Guinea	299		Near Oceania	299
Other Near Oceania	79		Other Near Oceania	79
Yap	151	Yap	Yap	0
Palau	117	Pala	Palau	0
Marianas	20	Mari	Marianas	0
Nauru	28	Naur	Nauru	0
Kosrae	21	Kosr	Kosrae	0
Kiribati	20	Kiri	Kiribati	0
Pohnpei	19	Pohn	Pohnpei	0
Marshall Islands	32		Marshall Islands	6
Kapingamarangi	33		Kapingamarangi	2
Fiji	10		Fiji	1
Vanuatu	73		Vanuatu	47
Tonga	63		Tonga	60
Samoa	52		Samoa	22
Marquesas	19		Marquesas	6
Other East Polynesia	19	Hawa, Rapa	Other East Polynesia	11
Cook Islands	90		Cook Islands	31
New Zealand	46		New Zealand	25
Total sequences	**1191**			**589**

Remote Oceania

Variation measures within and between regions were calculated using Arlequin. Levels of diversity within regions (θ_π Tables 6.5 and 6.6) were estimated from the infinite-site equilibrium relationship between θ and the mean number of pairwise differences (π) (where $E(\pi) = \theta$, Tajima 1983), with standard deviation calculated as the square root of the sampling variance described for gene diversity (Nei, 1987, Excoffier *et al.* 2005). Pairwise F_{ST} values were obtained, with significance testing using 1000 permutations. In Tables 6.5 and 6.6 the pairwise F_{ST} values that are significant at the 95% level and 99% level are boxed and shaded respectively.

The θ_π estimates range from 4.82 to 8.06 for the nt16065-nt16373 data set and from 0.686 to 5.174 for the nt16189-nt16370 Oceanic data set. New Guinea has the highest diversity in both data sets, but is within a standard deviation of all other groups. The 'Other Island Southeast Asia' category has the lowest diversity in the nt16065-nt16373 data set, followed by the sequences from the Americas. In the Oceanic nt16189-nt16370 data set the lowest θ_π estimates were from Kapingamarangi (a Polynesian outlier in Micronesia), where three haplotypes were found from 33 sequences.

Table 6.5 HVR-I nt16065-16373 data set diversity summary results
Shaded cells are significant (p<0.05)

	n	No. of haps.	θπ	S.D. θπ	Population pairwise FST												
					East Asia	Americas	Borneo	Japan	Taiwan	Southeast Asia	Other Is. SEA	Indonesia	Malaysia	Philippines	New Guinea	Other Near Oceania	Remote Oceania
East Asia	1285	575	7.12	3.70	-												
Americas	134	38	4.91	2.66	0.211	-											
Borneo	170	107	6.23	3.29	0.025	0.260	-										
Japan	476	235	6.60	3.45	0.017	0.210	0.036	-									
Taiwan	664	101	7.29	3.78	0.044	0.269	0.053	0.062	-								
Southeast Asia	197	158	7.23	3.76	0.014	0.268	0.031	0.034	0.042	-							
Other Is. SEA	15	4	4.82	2.79	0.060	0.326	0.097	0.077	0.099	0.052	-						
Indonesia	584	249	6.65	3.48	0.032	0.252	0.007	0.040	0.059	0.026	0.094	-					
Malaysia	61	47	7.00	3.70	0.004	0.255	0.017	0.014	0.044	0.004	0.065	0.017	-				
Philippines	146	73	6.86	3.59	0.028	0.257	0.014	0.038	0.024	0.034	0.093	0.024	0.024	-			
New Guinea	299	120	8.06	4.15	0.206	0.358	0.199	0.219	0.224	0.182	0.180	0.181	0.181	0.204	-		
Other Near Oc.	79	30	7.71	4.02	0.128	0.378	0.169	0.153	0.097	0.120	0.171	0.166	0.125	0.133	0.215	-	
Remote Oc.	211	84	7.37	3.83	0.137	0.361	0.166	0.166	0.094	0.133	0.183	0.165	0.137	0.128	0.219	0.011	-

Table 6.6 HVR-I haplotypes nt16189-16370 data set diversity summary results
Shaded and boxed cells are significant (p<0.05 and p<0.01)

	n	No. of haplotypes	θπ	S.D. θπ	Population pairwise FST																		
					New Guinea	Other Near Oceania	Yap	Palau	Marianas	Nauru	Kosrae	Kiribati	Pohnpei	Marshall Islands	Kapingamarangi	Fiji	Vanuatu	Tonga	Samoa	Marquesas	Other East Polynesia	Cook Islands	New Zealand
New Guinea	299	85	5.17	2.78	-																		
Other Near Oc.	79	27	4.9	2.67	0.21	-																	
Yap	151	13	2.86	1.68	0.42	0.12	-																
Palau	117	18	4.29	2.37	0.30	0.11	0.15	-															
Marianas	20	6	3.43	2.04	0.24	0.13	0.37	0.23	-														
Nauru	28	8	2.36	1.48	0.44	0.15	0.04	0.18	0.44	-													
Kosrae	21	6	3.28	1.96	0.41	0.13	0.09	0.18	0.35	0.07	-												
Kiribati	20	7	3.91	2.28	0.33	0.07	0.07	0.09	0.20	0.12	0.09	-											
Pohnpei	19	9	4.22	2.45	0.27	0.03	0.13	0.11	0.10	0.18	0.11	0.00	-										
Marshall Is.	32	9	2.9	1.74	0.41	0.14	0.06	0.16	0.36	0.07	-0.02	0.06	0.12	-									
Kapingamarangi	33	3	0.69	0.6	0.43	0.19	0.11	0.18	0.56	0.29	0.31	0.17	0.29	0.21	-								
Fiji	10	5	3.6	2.25	0.35	0.06	0.01	0.12	0.33	0.01	0.05	0.07	0.09	0.06	0.32	-							
Vanuatu	73	33	4.43	2.45	0.13	0.09	0.30	0.13	0.07	0.34	0.29	0.16	0.10	0.29	0.34	0.24	-						
Tonga	63	19	2.5	1.52	0.43	0.14	0.03	0.20	0.43	0.01	0.08	0.13	0.18	0.08	0.24	-0.03	0.34	-					
Samoa	52	14	2.02	1.28	0.46	0.19	0.06	0.24	0.52	0.02	0.12	0.21	0.27	0.12	0.33	0.01	0.40	0.00	-				
Marquesas	19	7	2.98	1.82	0.35	0.06	0.01	0.10	0.32	0.04	0.07	0.06	0.09	0.06	0.25	-0.02	0.22	0.01	0.07	-			
Other East Pol.	19	6	3.19	1.93	0.37	0.07	0.00	0.14	0.35	0.01	0.05	0.07	0.10	0.05	0.24	-0.05	0.26	-0.02	0.01	-0.02	-		
Cook Islands	90	14	3.59	2.04	0.36	0.08	0.04	0.16	0.34	0.05	0.08	0.10	0.12	0.08	0.17	-0.03	0.27	0.02	0.04	0.01	-0.02	-	
New Zealand	46	11	2.12	1.33	0.43	0.15	0.02	0.18	0.46	0.01	0.09	0.14	0.21	0.08	0.26	-0.01	0.34	-0.01	0.01	0.00	-0.01	0.03	-

All of the pairwise comparisons for the nt16065-nt16373 data set except those between Malaysia and East Asia, Indonesia and Southeast Asia, and Remote Oceania and the 'Other Near Oceania' (not New Guinea) category showed significant differentiation between regions ($p<0.01$). The highest pairwise F_{ST} values were between the Americas and other regions, particularly Oceania. The pairwise F_{ST} values for the Oceanic nt16189-nt16370 data set also show differences between the regions defined, with all comparisons for New Guinea, Kapingamarangi and Vanuatu significant ($p<0.01$). The comparisons between New Guinea and other regions have the highest values. The Polynesian regions (Tonga, Samoa, Marquesas, Other East Polynesia, Cook Islands and New Zealand) by contrast have low pairwise F_{ST} values, with only two values (for Samoa and the Marquesas, and Samoa and the Cook Islands) significant at the 95% level.

In the larger nt16065-nt16373 data set containing sequences from Oceania, Asia and the Americas the original 4321 sequences were reduced to 1544 haplotypes, of which a large proportion (1041, approximately 67%) were found in single individuals. Of the 503 haplotypes represented by two or more sequences 166 were shared between regions: 109 were found in 2 regions, 30 in three regions and 14 in four regions. The remaining 13 haplotypes were found in more than five regions, and are described in Table 6.7. A total of 53 individuals had the distinctive pre-Polynesian motif haplotype (EF077392NZTong), which is one of the most widespread, found in seven of the 13 regions within the data set.

The 16129A 16172C 16304C motif appears four times (U47181Phil1, EF06897Phil+16311C, DQ309863Kark+16172C 16362C and EF069246Wain+16162G) in Table 6.8, and variations of 16223T 16362C are also common (DQ309865Kark, AB093907Mala+16311C, DQ372876Mars+16295C and DQ309860Kark+16291T). While the distributions of these haplotypes may reflect recent population movements it is also possible that some of these sequences are convergent, rather than identical by shared descent. The sites involved demonstrate high levels of homoplasy, with four of the five bases in the motifs requiring on average more than 4 mutation events to fit the 75-taxa coding-region trees in the homoplasy analysis described in Chapter Five (Figure 5.2).

Details of the 221 haplotypes from the nt16065-nt16373 data set which included samples from Oceania have been tabulated and are included as Appendix E.6.1. This table summarises the frequencies and distributions of the haplotypes, and lists their differences to the rCRS. 43 haplotypes included entire mt genome sequences and these are indicated with their haplogroup details. Where possible the HVR-I haplotypes have been assigned to haplogroups (68 could not be identified with any confidence). More than half (126) of the Oceanic haplotypes were represented by single individuals. Of the 95 haplotypes found in more than one individual 25 were also present in regions outside of Oceania (Appendix E.6.1).

Table 6.7 HVR-I nt16065-16373 data set haplotypes found in five or more regions

Haplotype	n	Regions	Differences to rCRS nt16065-nt16373
AB093865Mala	17	East Asia, Japan, Southeast Asia, Indonesia, Malaysia	16189C 16223T 16362C
DQ149036Thai	11	East Asia, Borneo, Japan, Taiwan, Southeast Asia	16223T
EF068973Phil	21	Borneo, Taiwan, Indonesia, Malaysia, Philippines	16129A 16172C 16304C 16311C
U47251Phil1	24	East Asia, Borneo, Taiwan, Indonesia, Philipines	16157C 16256T 16304C 16335G
AB093907Mala	18	East Asia, Borneo, Japan, Southeast Asia, Indonesia, Malaysia	16223T 16311C 16362C
DQ309863Kark	34	East Asia, Borneo, Taiwan, Indonesia, Philippines, New Guinea	16129A 16172C 16294T 16304C 16362C
DQ372876Mars	79	Borneo, Taiwan, Indonesia, Malaysia, Philippines, Remote Oceania	16223T 16295T 16362C
EF069246Wain	37	East Asia, Borneo, Japan, Taiwan, Southeast Asia, Indonesia	16129A 16162G 16172C 16304C
U47181Phil1	29	East Asia, Borneo, Southeast Asia, Indonesia, Malaysia, Philippines	16129A 16172C 16304C
AB119340Balo	10	Borneo, Taiwan, Indonesia, Malaysia, Philippines, New Guinea, Other Near Oceania	16126C 16129A 16223T 16297C
DQ309865Kark	104	East Asia, Borneo, Japan, Taiwan, Indonesia, Philippines, New Guinea	16223T 16362C
EF077392NZTo	53	East Asia, Taiwan, Indonesia, Malaysia, Philippines, New Guinea, Remote Oceania	16182CtvA 16183CtvA 16189C 16217C 16261T
DQ309860Kark	93	East Asia, Borneo, Japan, Taiwan, Indonesia, Malaysia, Philippines, New Guinea	16223T 16291T 16362C

The nt16189-nt16370 data set of 1191 sequences from Oceania consisted of 199 haplotypes, 97 of which were found in single individuals. This data set has greater coverage of Remote Oceania than the longer data set (Table 6.4) and was used for the detailed analyses of HVR-I haplogroup phylogenies and distributions described below.

124 of the 199 haplotypes were assigned to haplogroups using the entire mt genome phylogenies (Chapter Three) as a guide. Polymorphisms between nt16189 and nt16370 were identified in the haplogroups present in Oceania. Those which occurred in the higher branches of the trees, and preferably showed relatively low rates of homoplasy in the analysis in Chapter Five (Figure 5.2) were selected as defining polymorphisms for haplogroups (Table 6.8). The M/M7c haplogroup, with defining polymorphisms 16223T 16295T, is perhaps

Table 6.8 HVR-I haplogroup defining polymorphisms (nt16189-nt16370 data set)

Haplogroup	Defining polymorphisms	Haplotypes (n)	Sequences (n)
N/R/B4a	16189C 16217C	42	620
M/Q1	16265CtvA 16343G	34	183
N/R/P1	16266T 16357C	23	102
M/M7c	16223T 16295T	4	23
M/M27a	16189C 16223T 16311C 16320	2	5
M/M27c	16223T 16301T 16304C	1	2
M/M28b	16223T 16318TtvA	2	2
M/Q not Q1	16241G	16	59
Unassigned		75	195
Total		**199**	**1191**

the least reliable of these categories as 16223T shows high levels of mutability and there is only a single M/M7c entire mt genome sequence from Oceania for reference (DQ372876, Figure 3.5). 75 haplotypes, representing 195 sequences (16% of the data set), could not be assigned to a haplogroup. Appendix E6.2 lists all haplotypes within the nt16189-nt16370 data set by haplogroup assignment with their polymorphisms relative to the rCRS and frequencies in the regions defined.

The distribution of the three main Oceanic haplogroups is shown in Figure 6.1. The 'young' N/R/B4a haplogroup is the most common lineage in Remote Oceania, reaching high frequencies in Polynesia and Micronesia, and accounting for more than half (620/1191) of all samples in the nt16189-nt16370 data set. In Near Oceania N/R/B4a haplotypes are found in only 4% of the New Guinea samples, but make up almost half (48%) of the samples in the 'Other Near Oceania' group.

A large proportion of the samples (59 of the total 79) contributing to the 'Other Near Oceania' group come from an Austronesian-speaking community in the Balopa Islands of the Manus province of Papua New Guinea, located at the north-western end of the Bismarck Archipelago (Ohashi *et al.* 2006). Friedlaender *et al.* (2007) have constructed a comprehensive data set of HVR-I and HVR-II sequences from Northern Melanesian populations (n=1223, from 32 populations in New Ireland, New Britain, Bougainville and Malaita). They found N/R/B4a haplotypes reached high frequencies in New Ireland, Bougainville and Malaita (in some cases in non-Austronesian speaking populations), but were relatively rare in populations from New Britain.

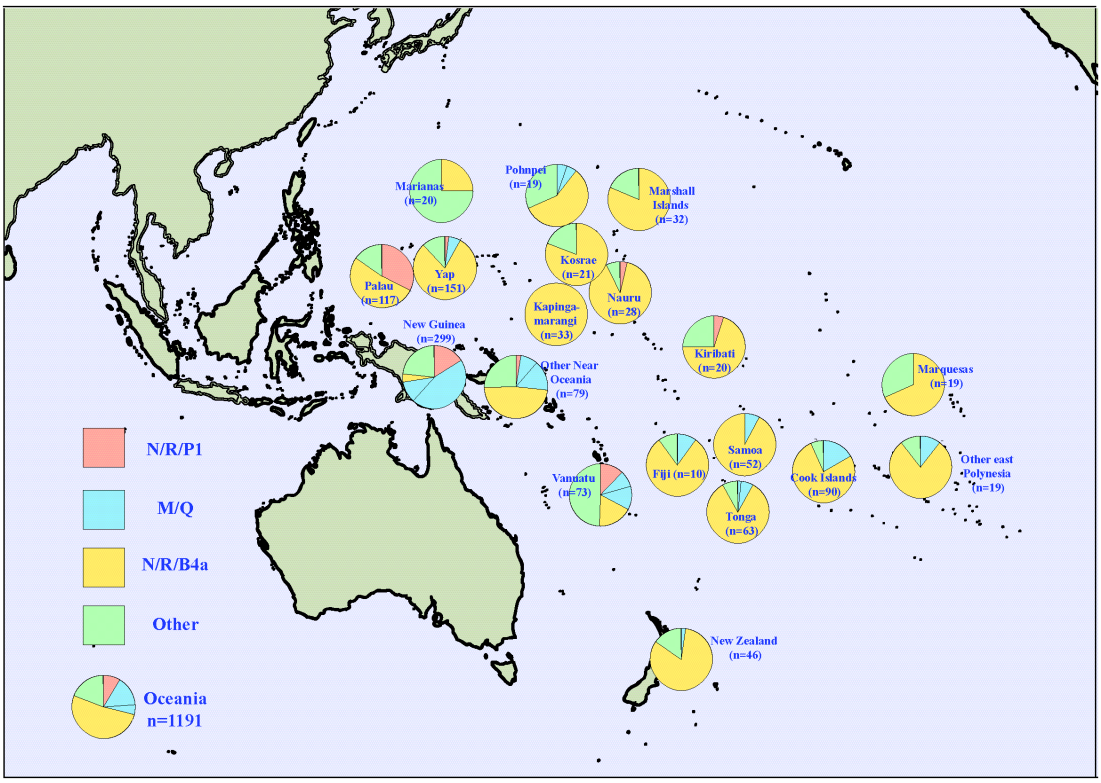


Figure 6.1 Distribution of HVR-I haplotypes (nt16189-nt16370) in Oceania
The distribution of three common haplogroups found in Oceanic populations. The haplotypes were assigned to haplogroup categories as described in Table 6.8. In this diagram for simplicity the M/Q1 and 'M/Q not Q1' have been coloured to a single M/Q category (in blue) but a dividing line remains where types from both groups were present in the population.

Sequences belonging to the 'ancient' N/R/P1 and M/Q haplogroups are also found in both Near and Remote Oceania, at varying frequencies (Figure 6.1). N/R/P1 sequences are common in New Guinea, and are found in populations from Micronesia (Palau, Yap, Nauru and Kiribati) and Vanuatu but not in Polynesian samples. In Figure 6.1, haplotypes assigned to the M/Q1 and M/Q not Q1 categories defined in Table 6.8 are grouped together as M/Q. M/Q haplotypes are the most common in the New Guinea sample, and are present in Vanuatu and Polynesia, but are not found in the eastern Micronesian island groups.

In the following section, the sequences from the nt16189-nt16370 data set assigned to N/R/B4a, N/R/P1 and M/Q1 haplogroups are described in greater detail, and links to populations outside of Oceania explored through analysis of the nt16065-nt16373 sequence set.

6.4 Phylogenies from HVR-I sequences: N/R/P1, M/Q1 and N/R/B4a

Phylogenetic analyses of human mtDNA control region sequences can be difficult due to two seemingly contradictory factors: firstly, the time elapsed from common ancestry is often very short in an evolutionary sense resulting in low levels of variability and a lack of resolution of the lineage history, and secondly, high and unequal rates of substitutions at certain sites can result in parallel, or convergent polymorphisms in different lineages, or multiple substitutions at a single site obscuring patterns of shared descent. Consequently it is unusual to find a single tree from control region data which can best explain the history of the sequences, and networks are often highly complex.

These issues were examined in Chapter Five with the aim of determining a system of weighting for the HVR-I sites which would enable the existing collections of sequences from Oceania to be analysed with greater confidence. As the weighting schemes devised did not show any significant improvement over the use of unweighted characters the analysis of the three major haplogroups detailed below was undertaken conservatively, with parsimony analyses and the phylogenies from entire mt genomes used to guide the reconstruction of the trees.

N/R/P1

The N/R/P1 subset of the nt16189-nt16370 data set consisted of 102 samples belonging to 23 haplotypes, defined by transitions at nt16266 and nt16357 (Figure 6.2a). There were 10 informative sites in the 180 characters of the data set, and a PAUP* search (version 4. 0b10, Swofford 2003) found a single tree with a score of 14, re-drawn in Figure 6.2b. The 16266T 16270T 16357C haplotype has the highest frequency, with 51 samples having this type from Vanuatu, New Guinea and Yap, and several haplotypes branch off with additional substitutions from this haplotype. The 16270T variant occurs in two of the P1 entire genome sequences (AF347005 and AY289097 from highland Papua New Guinea, at the right of Figure 6.2a), which also share two other changes in the HVR-I region analysed (16257T 16354T). As these two sequences are separated by four and two coding region changes respectively to their common ancestor, and the other HVR-I polymorphisms they share are not seen in most HVR-I sequences with nt16270T, it is possible that this variant occurred early in the development of the P1 lineage.

A transversion at nt16259 separates three haplotypes (AF066301, AF066226, AF066120) found in Micronesian samples from the U4721 haplotype (Figure 6.2). The samples come from Yap, Nauru and

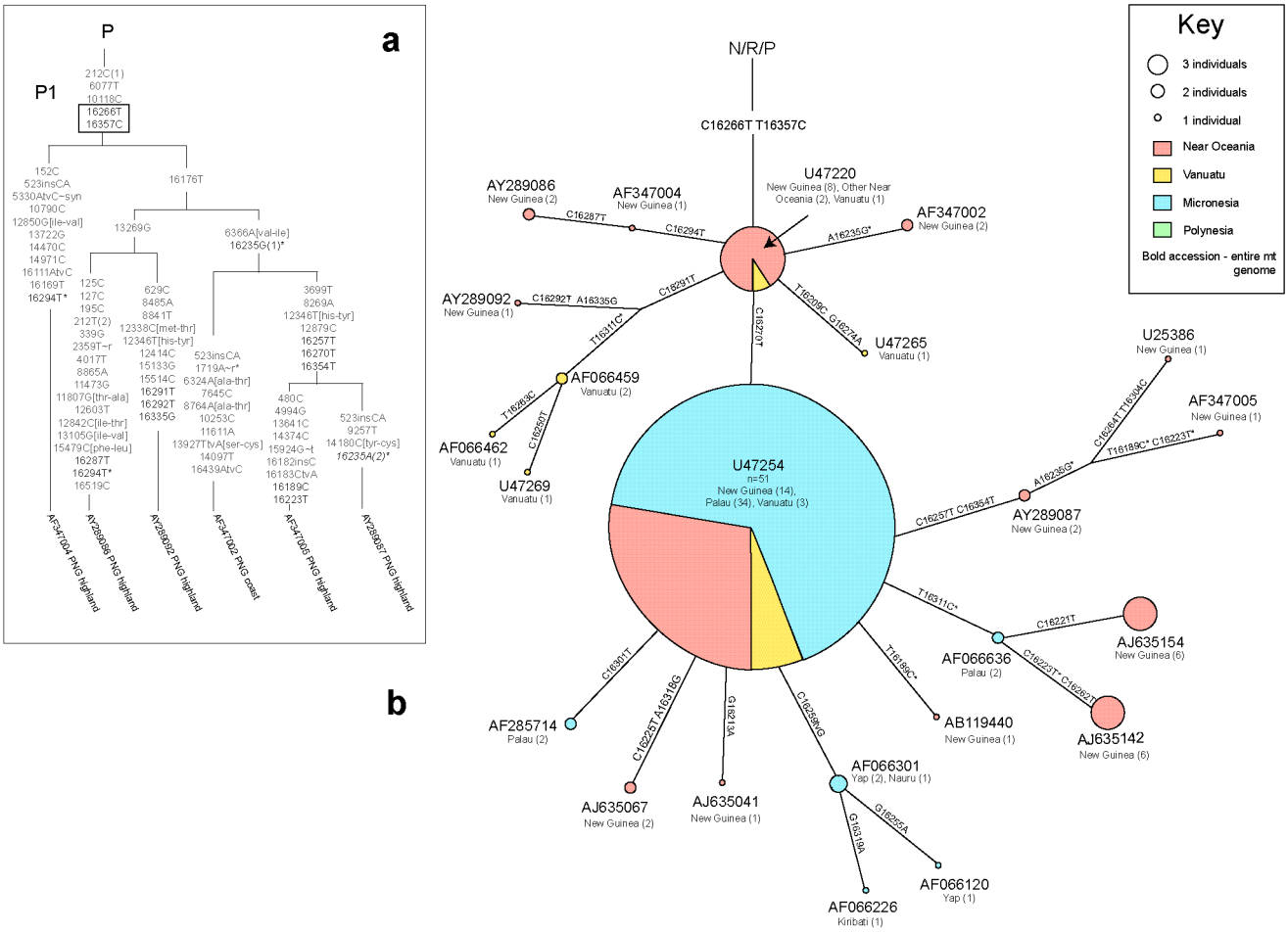


Figure 6.2 N/R/P1 HVR-I haplotypes in Oceania
a) Labelled phylogeny of six entire N/R/P1 mt genomes. The polymorphisms used to define P1 haplotypes from HVR-I sequences are boxed, and other HVR-I variants between nt16189-nt16379 shown in black type. b) Tree of HVR-I haplotypes, with vertex size reflecting the number of sequences within each haplotype. Haplotypes named in bold type are from whole genome sequences, and bases requiring more than one step in the tree are marked with an asterisk.

Kiribati, and this polymorphism may prove useful for tracing movements from Near Oceania to the Micronesian island groups, and within Micronesia.

The nt16065-nt16373 HVR-I data set contained 24 P1 haplotypes, representing 54 samples all of which came from Near Oceania. This data set was searched for any haplotypes other than those from Oceania with the N/R/P1 polymorphisms 16266T and 16357C using MacClade (version 4.06, Sinauer Associates Inc., Sunderland, Massachusetts). Two additional haplotypes were identified with these variants, both from single samples. The first haplotype, EF068809Mana (16176T 16221T 16266T 16325C 16357C), was found in two samples from Manado in Sulawesi (Hill *et al.* 2007).

The second, AP008719Japa, has the HVR-I profile 16093C 16129A 16223T 16266T 16311C 16357C

and is derived from an entire genome sequence from Japan. The occurrence of the 16266T and 16357C polymorphisms in this haplotype are likely to be convergent rather than reflecting common ancestry as the haplotype tree for the global data set (Appendix D4.1) places this sequence within macrohaplogroup M.

M/Q1

The Q1 subset of the nt16189-nt16370 data set was assigned by a transversion from A to C at nt16265 and a transition to G at nt16343. There were 34 haplotypes within the data set with these polymorphisms, representing 183 sequences (Figure 6.3). The majority of the Q1 samples from Polynesia (n=20) belong to either the most common Q1 haplotype which has no additional HVR-I mutations (AY289090) or a haplotype with one additional transition at nt16293 (DQ372884). The small number of Micronesian sequences (n=10) share a single haplotype with samples from Vanuatu and New Guinea (U47169).

Four haplotypes from the nt16065-nt16373 data set were found to be shared with the Oceanic Q1 samples. Four samples from Indonesia had the Q1 haplotype (16223T 16241G A16265tvC 16311C 16343G), and six had this sequence with the rCRS state at 16223C. Nine individuals from Island Southeast Asia had the core Q1 haplotype + 16270T.

The HVR-I tree (main diagram, Figure 6.3) appears to document a rapid expansion from an ancestral haplotype represented by the AY289090 haplotype. However, the entire mt genomes within the data set (Figure 6.3a) have a number of coding region changes along the terminal branches which suggest a longer period to common ancestry than indicated by the HVR-I phylogeny.

N/R/B4a

The most common haplogroup found in the Oceanic HVR-I data set (nt160189-nt16370) was N/R/B4a, with 620 samples belonging to 40 haplotypes from this haplogroup (defined by 16189C, 16217T). The tree reconstruction shown in Figure 6.4 follows the whole genome phylogeny (Figure 3.4), with dotted branches reflecting the uncertainties seen in the consensus network of parsimony trees generated, and recurrent changes on the tree marked with an asterisk. There were 10 informative sites in the subset of 40 haplotypes, and heuristic parsimony searches found 11080 trees with score 18 (not shown).

The whole genome phylogeny (Figure 3.4) has shown that the nt16247G transition has occurred relatively recently in prehistory and the haplotype frequencies shown in the HVR-I tree (Figure 6.4) appear to reflect a recent expansion from a common ancestral type. To assess the use of the HVR-I rho dating technique discussed in Chapter Three on a data set much closer to a population sample than the comparatively sparse

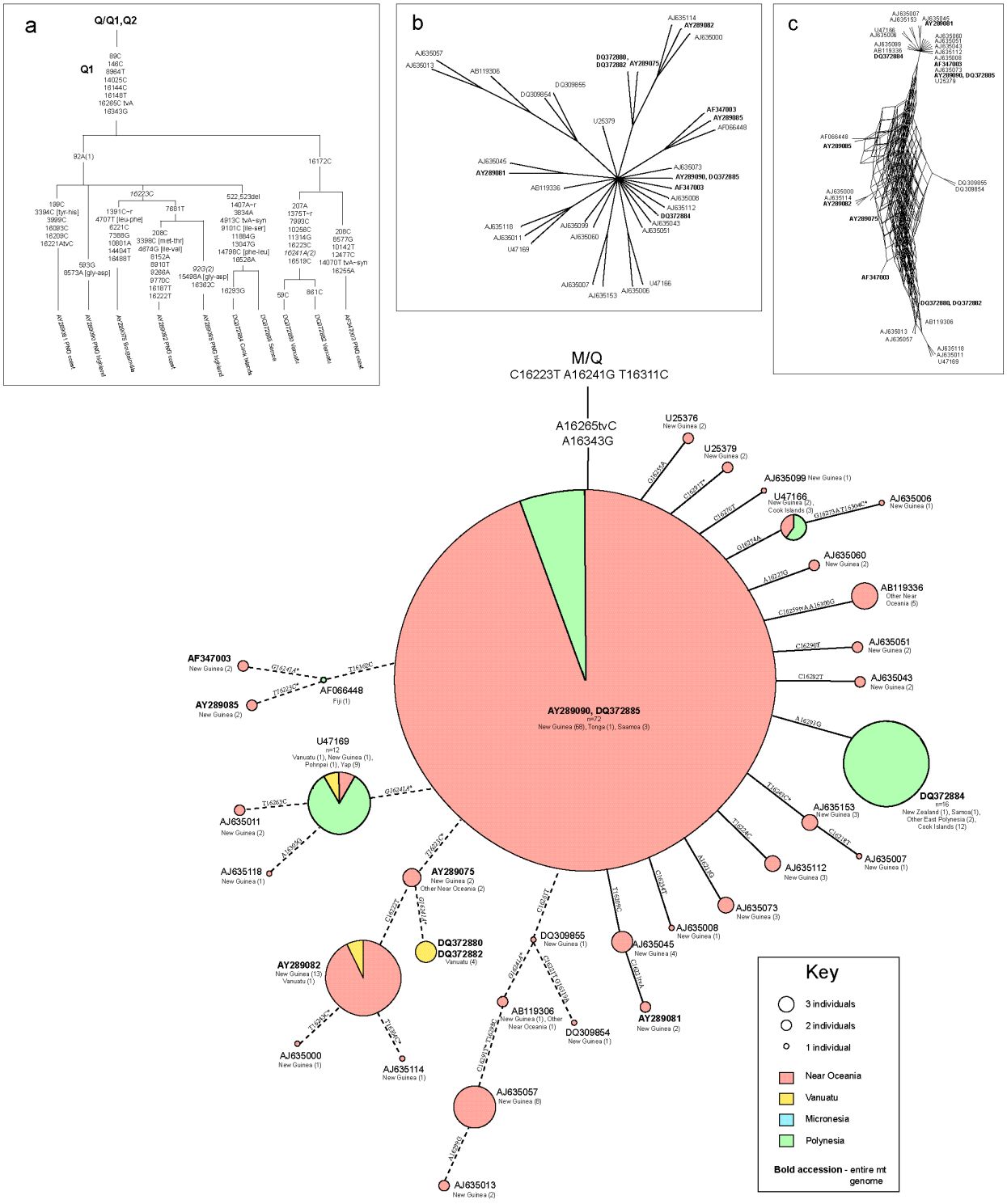


Figure 6.3 M/Q1 HVR-I haplotypes in Oceania
Insets: a) Labelled phylogeny of ten entire M/Q1 mt genomes b) The first tree found by parsimony search (34 haplotypes, 11 parsimony informative characters, heuristic search score 18). c) The consensus network of 494 most parsimonious trees found. Main figure: Labelled tree of 183 HVR-I samples belonging to 34 haplotypes, reconstructed from the first tree found (inset b). The dotted branches have the least support as shown by the consensus network (inset c). The vertex size reflects the number of sequences within each haplotype. Haplotypes named in bold type are from whole genome sequences, and bases requiring more than one step in the tree are marked with an asterisk.

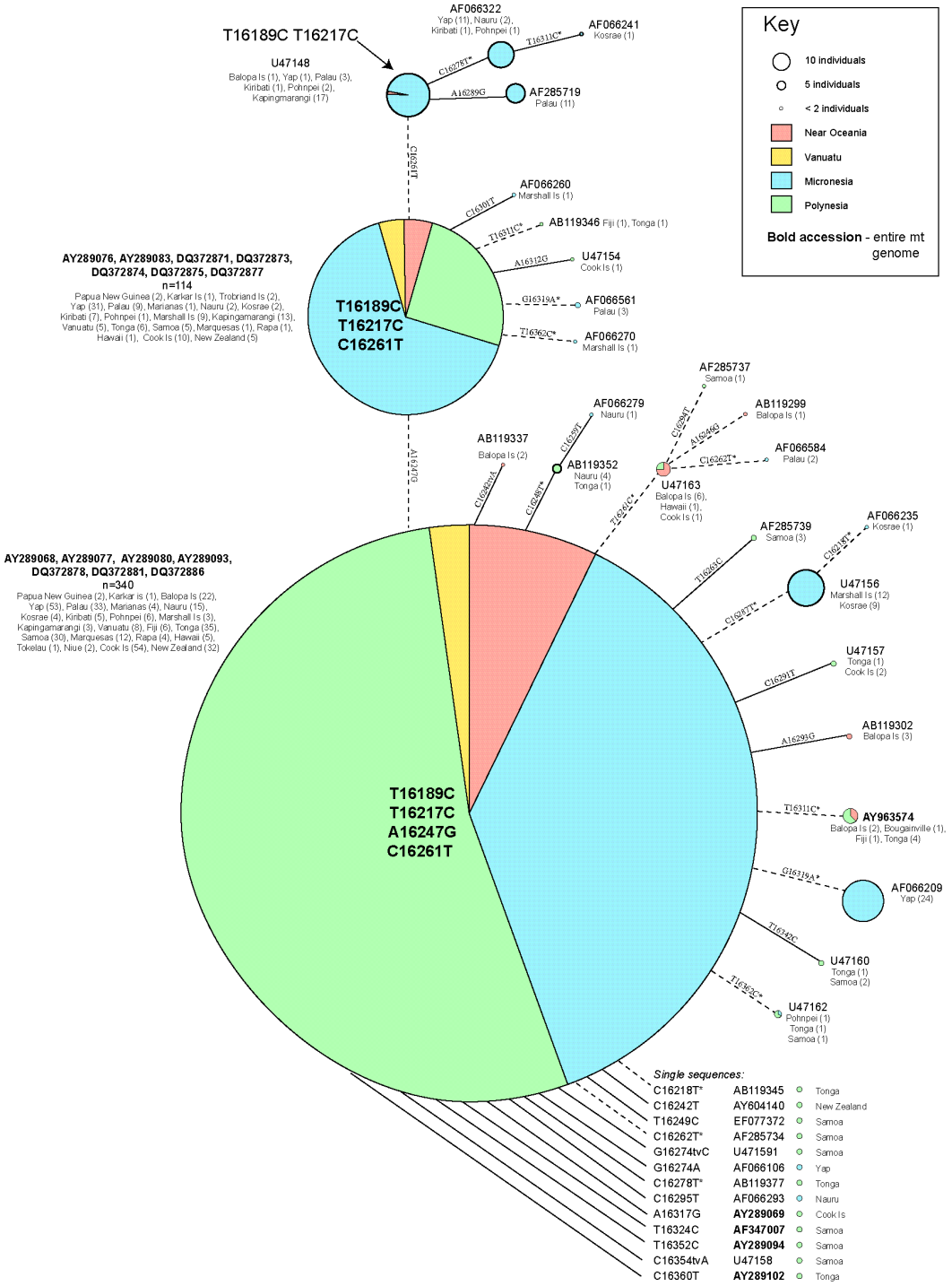


Figure 6.4 N/R/B4a HVR-I haplotypes from Oceania
Labelled tree of 40 haplotypes (n=620 samples) from Oceanic populations. The dotted branches have the least support as shown by the consensus network of the most parsimonious trees found. The vertex size reflects the number of sequences within each haplotype. Haplotypes named in bold type are from whole genome sequences, and bases requiring more than one step in the tree are marked with an asterisk.

and non-randomly sampled whole mt genomes, the HVR-I N/R/B4a samples with the 16247G variant (B4a1a1PM) were dated using the Network program (version 4.2.01, Fluxus Technology Limited, ©2004-2006). The input file for the 29 haplotypes (n=446 sequences) was created using DnaSP (version 4.10.9, Rozas *et al.* 2003), and the age to the central 16247G haplotype measured using the default mutation rate of 20180 years per mutation. This gave an age of the 16247G haplotype of 6932 years with a standard deviation of 1868 years. The average rate of change estimated from the ratio of HVR-I to synonymous coding region changes in the entire mt genome phylogenies described in Chapter Three (1 change per 10 569 years) was also used to estimate the date of the 16247G vertex using Network. This resulting date was 3631 years, with a standard deviation of 978 years.

Five haplotypes were also found outside of Oceania. The 16189C 16217C haplotype (sometimes known as the pre-pre-Polynesian motif') was found in six samples within the nt16065-nt16373 data set, in individuals from East Asia (Buriat and Mien) and in four samples from Peru. The 16189C 16217C 16261T (pre-Polynesian motif) was found in five East Asian samples, 27 Taiwanese, three Philippines samples and one sample from both Indonesia and Malaysia. Haplotypes with the pre-Polynesian motif plus one variant (16093C, 16129A or 16311C) were found in 20 East Asian and 10 Taiwanese samples.

6.3 Discussion

The distribution of HVR-I haplotypes in Remote Oceania shows different levels of contribution of the 'ancient' Near Oceanic lineages to the present-day populations; with particularly marked differences seen between the diversity found in Vanuatu, and the homogeneity of Polynesia. The Micronesian islands also show varying levels of the ancient P and Q haplotypes, with the larger proportion of haplotypes unassigned to type in the west perhaps reflecting early settlement from Island Southeast Asia (Lum and Cann 2000).

The unassigned haplotypes from Polynesia (Appendix E6.2) may derive from historic period migration (Whyte *et al.* 2005, for example AY604130NZMaor, AY604136NZMaor), but it is difficult to confirm this without additional testing of coding region polymorphisms. The haplotype AY604133NZMaor, for example, shows no differences from the reference sequence and is present in 13 individuals from New Guinea (n=8), Vanuatu (n=4) and New Zealand (n=1). The entire mt genome analysis in Chapter Three revealed this haplotype to be present in N/R/P2 samples (Figure 3.2), and the comparison of coding and HVR-I phylogenies in the previous chapter highlighted the difficulties in distinguishing between the N/R/P and N/R/

HV haplogroups on HVR-I sequence data alone.

The newly reported M/M9/E entire genomes from Near Oceania (Friedlaender *et al.* 2007) raise the question of whether these haplotypes are present in Remote Oceania as well. Unfortunately there are few polymorphisms in the nt16189-nt16370 region to identify this haplogroup (Friedlaender 2007:7), and without further analysis the presence of M9/E haplotypes cannot be confirmed from the data set.

The sample sizes throughout Oceania remain small, particularly in Remote Oceania. This could be improved by the incorporation of existing sequence sets (for example Cox 2003, Friedlaender *et al.* 2006) into the data set described here. The collection of more samples (and analysis of diagnostic coding polymorphisms in combination with control region sequencing) would also be of great benefit in addressing issues such as the number and likely geographic source of Remote Oceanic founding lineages.

.

7. Conclusions and Future Directions

The collection and analysis of the whole genome sequences described in this thesis provides a solid framework for interpreting the relationships between the mitochondrial (mt) haplogroups present in Oceania. The distribution of haplotypes forms a clear division between the 'ancient' haplogroups N/R/P, M/Q, M/M27, M/M28, and M/M29 found in Oceania and those geographically restricted to Australia (N/S, N/O, R/R12, M/M42); and the 'young' haplogroups N/R/B4a and M/M7bc which share common ancestry relatively recently with samples from outside of Oceania. Recent work by Friedlander *et al.* (2006) has revealed haplotypes from a third 'young' haplogroup, M/M9/E also present in Near Oceania.

The 'ancient' haplogroups of Oceania coalesce at the 'Out-of-Africa' L3 polytomy and reflect the deep human history in these regions, where early settlement dates to at least 40 000BP (O'Connell and Allen 2004). In contrast, the phylogenies of the 'young' haplogroups, N/R/B4a and M/M7bc, reveal relatively recent common ancestry between the Oceanic haplotypes and others from Asia, although the number of samples are limited, particularly for M/M7bc. The Oceanic samples in N/R/B4a form a subgroup, B4a1a, of the tree which also includes Taiwanese sequences, while the single Micronesian M/M7bc sample belongs to the M7bc/c subgroup and is most closely related to samples from Mongolia, the Philippines and Taiwan. The TMRCA date estimates from synonymous coding changes (Table 3.2) for the B4a1a subgroup, containing Taiwanese and Oceanic samples, and the B4a1a1 vertex which contains only Oceanic samples are 10 822 ± 1759 and 5276 ± 1429 years respectively, indicating a recent entry to Oceania of the ancestral haplotypes from this haplogroup, broadly coinciding with the appearance in the archaeological record of Lapita settlements in Near Oceania from 3400BP (Kirch 2000).

The distribution of control region haplotypes in Oceania reviewed in Chapter Six demonstrates that a subset of the mtDNA variation present in Near Oceania has been carried out into Remote Oceania. The N/R/B4a haplotypes predominate in Remote Oceania, with low frequencies of 'ancient' M/Q haplotypes seen in Polynesia, and varying frequencies of M/Q and N/P present in Island Melanesia and Micronesia (Figure 6.1). If the contemporary distributions reflect founding populations, the mixture of 'ancient' and 'young' haplotypes in Remote Oceanic populations indicate integration in Near Oceanic prehistory between the incoming peoples carrying the 'young' haplotypes and those already present: providing support for Green's (1991) third and fourth model sets. At present inadequate whole mtDNA sampling from Island Southeast Asia leaves the origin of the immigrants to Oceania carrying the 'young' haplotypes in question, although close relationships are seen between the N/R/B4a1a Oceanic haplotypes and Taiwan.

During the course of this project the number of human mt genomes available on public databases has grown dramatically, and with improvements in sequencing technologies this is likely to continue. It is relatively straightforward to add new sequences from the region into the existing large data set of all Oceanic haplotypes, as was done in Chapter Two for the recently described Australian sequences (van Holst Pellekaan *et al.* 2006). It is anticipated an increase in mitochondrial genome sequences from haplogroups at present under-represented in the data set (for example R/R12, R/R21, N/P/P7, M/M21, M/M22 and N/N22) will aid in clarifying the branching order from the major macrohaplogroup vertices.

The detailed phylogenies constructed in Chapter Three include the first M/Q entire genome sequences described from Remote Oceania. There are distinct subtypes seen in the samples from Polynesia and Vanuatu (Figure 3.1), which are quite distant from the other sequences from Near Oceania. Future analyses could be targeted towards finding the nearest relatives of these sequence types in Near Oceania, using the single nucleotide polymorphisms (SNPs) unique to each of the two subtypes. SNPs in the N/R/B4a1a phylogeny were used in this study to supplement control region sequencing of the sample set from Auckland (Chapter Six), illustrating the power of this approach, and advantages over entire mt genome sequencing.

Two N/R/P haplotypes obtained from this study, DQ372870 and DQ372872, from the Trobriand Islands in Near Oceania, clearly fall within the N/R/P2 haplogroup (Figure 3.2), yet share an HVR-I haplotype with the European N/R/HV reference sequence. This coincidence inspired further work over the course of this project examining the incidence of recurrent mutations in mtDNA, and the implications high rates of homoplasy could have for phylogenies derived from partial sequences.

The global data sets assembled over the course of this study, described in Chapter Four, are a resource for future analyses of human mtDNA phylogenies, and the properties of the molecule itself, and it is my intention to maintain these by the inclusion of new sequences as they become available. At present the L3/N haplotypes are over-represented in the data set (Table 4.3), but it is hoped future analyses may help to adjust the balance; for example a recent study (Gonder *et al.* 2007) describes 62 new African sequences.

The assessment of variation and tests of selections undertaken in Chapter Four highlighted diversity in the *ATP6* gene, whose protein product contributes to Complex V of the OXPHOS system. The McDonald-Kreitman test also found significant evidence for non-neutral evolution in two genes, COIII and Cyt *b*. A phylogenetic approach, using the large number of sequences now available, may in future contribute to our understanding of the roles positive and negative selection have played in the history of human mtDNA.

The positive results for recombination from the phi test (Bruen *et al.* 2006) of subsets of the global data sets (Chapter Four) are intriguing, and a further indication of the high levels of homoplasy present in human mtDNA seen in the phylogenetic reconstructions of Oceanic sequences, and the analyses in Chapter Five. Determining the most likely cause of these unexpectedly high levels of homoplasy, recombination or extreme mutation rate variability, will be an interesting and challenging topic for future research.

The phylogenetic analysis of control region homoplasy undertaken in Chapter Five provides a relative scale of mutability in this region (Figure 5.2), supporting previous findings of 'hypervariability' at certain positions (Forster *et al.* 2002, Meyer *et al.* 1999, Stoneking 2000). The mtDNA coding region was used to construct the trees on which the control sequence homoplasy was assessed, and the analysis of the parsimony length of coding characters on these trees also gave indications of increased mutability at certain positions (Table 5.1). The measures of homoplasy generated by this analysis allow the unusual patterns, such as those seen in the N/R/P haplogroup, to be investigated further.

While this project had the stated aim of investigating mtDNA diversity in contemporary Oceanic populations to infer patterns of population movements in prehistory, the intriguing patterns emerging from the analyses led to an equal focus on the properties and microevolution of the mtDNA molecule itself. The results of analyses undertaken in this project in both of these areas have raised many questions still to be addressed, such as the issues of homoplasy and recombination outlined above. In the context of Oceanic prehistory the results presented here document the presence in Oceania of two distinct subsets of mtDNA haplogroups; designated 'ancient' and 'young'. The distribution of these types throughout Remote Oceania is suggestive of small founding populations, and different, although perhaps overlapping, sources for the migrants to Polynesia, Vanuatu and the islands of west Micronesia.

The issues involved in assigning an absolute to date to vertices in mtDNA phylogenies in the recent past (Chapter Three) lead me to conclude that while molecular dating is unlikely to contribute greatly to the understanding of the timing of the colonisation events in Remote Oceania, it will be of increasing value to the interpretations of the much older mtDNA histories from Near Oceania. The collection of new samples from throughout Remote Oceania, and analysis with reference to the entire mt genome phylogenies using a combination of sequencing and other techniques (for example SNaPshot minisequencing, Quintans *et al.* 2004), is likely to make the biggest impact on our understanding of the prehistory of this region, allowing a more detailed picture of the geographic source of the settlers, and post-settlement population histories and interactions to be reconstructed.

REFERENCES

Abu-Amero, K. K., A. M. Gonzalez, J. M. Larruga, T. M. Bosley, and V. M. Cabrera. 2007. Eurasian and African mitochondrial DNA influences in the Saudi Arabian population. BMC Evolutionary Biology 7:32.

Achilli, A., C. Rengo, V. Battaglia, M. Pala, A. Olivieri, S. Fornarino, C. Magri, R. Scozzari, N. Babudri, A. S. Santachiara-Benerecetti, H. J. Bandelt, O. Semino, and A. Torroni. 2005. Saami and Berbers - An unexpected mitochondrial DNA link. American Journal of Human Genetics 76:883-886.

Achilli, A., C. Rengo, C. Magri, V. Battaglia, A. Olivieri, R. Scozzari, F. Cruciani, M. Zeviani, E. Briem, V. Carelli, P. Moral, J. M. Dugoujon, U. Roostalu, E. L. Loogvali, T. Kivisild, H. J. Bandelt, M. Richards, R. Villems, A. S. Santachiari-Benerecetti, O. Semino and A. Torroni. 2004. The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. American Journal of Human Genetics 75:910-918.

Amo, T. and M. D. Brand. 2007. Were inefficient mitochondrial haplogroups selected during migrations of modern humans? A test using modular kinetic analysis of coupling in mitochondria from cybrid cell lines. Biochemical Journal 404:345-351.

Anderson, S., A. T. Bankier, B. G. Barrell, M. H. de Bruijn, A. R. Coulson, J. Drouin, I. C. Eperon, D. P. Nierlich, B. A. Roe, F. Sanger, P. H. Schreier, A. J. Smith, R. Staden, and I. G. Young. 1981. Sequence and organization of the human mitochondrial genome. Nature 290:457-465.

Andrews, R. M., I. Kubacka, P. F. Chinnery, R. N. Lightowlers, D. M. Turnbull, and N. Howell. 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nature Genetics 23:147.

Aquadro, C. F., and B. D. Greenberg. 1983. Human mitochondrial DNA variation and evolution: analysis of nucleotide sequences from seven individuals. Genetics 103:287-312.

Arnason, U., X. Xu and A. Gullberg. 1996. Comparison between the complete mitochondrial DNA sequences of Homo and the common chimpanzee based on nonchimeric sequences. Journal of Molecular Evolution 42:145-152.

Atkinson, Q. D. 2006. From Species to Languages - A phylogenetic approach to human prehistory. PhD thesis, University of Auckland.

Awadalla, P., A. Eyre-Walker, and J. M. Smith. 1999. Linkage disequilibrium and recombination in hominid mitochondrial DNA. Science 286:2524-2525.

Ballinger, S. W., T. G. Schurr, A. Torroni, Y. Y. Gan, J. A. Hodge, K. Hassan, K. H. Chen, and D. C. Wallace. 1992. Southeast Asian mitochondrial DNA analysis reveals genetic continuity of ancient mongoloid migrations. Genetics 130:139-152.

Bandelt, H. J., P. Forster, and A. Rohl. 1999. Median-joining networks for inferring intraspecific phylogenies. Molecular Biology and Evolution 16:37-48.

Bandelt, H. J., P. Forster, B. C. Sykes, and M. B. Richards. 1995. Mitochondrial portraits of human populations using median networks. Genetics 141:743-753.

Bandelt, H. J., C. Herrnstadt, Y. G. Yao, Q. P. Kong, T. Kivisild, C. Rengo, R. Scozzari, M. Richards, R. Villems,

V. Macaulay, N. Howell, A. Torroni, and Y. P. Zhang. 2003. Identification of native American founder mtDNAs through the analysis of complete mtDNA sequences: Some caveats. Annals of Human Genetics 67:512-524.

Bandelt, H. J., and T. Kivisild. 2006. Quality assessment of DNA sequence data: autopsy of a mis-sequenced mtDNA population sample. Annals of Human Genetics 70:314-326.

Bandelt, H. J., Q. P. Kong, W. Parson, and A. Salas. 2005. More evidence for non-maternal inheritance of mitochondrial DNA? J Med Genet 42:957-960.

Bandelt, H. J., V. Macaulay, and M. Richards. 2000. Median networks: speedy construction and greedy reduction, one simulation, and two case studies from human mtDNA. Molecular Phylogeny and Evolution 16:8-28.

Bandelt, H. J., A. Olivieri, C. Bravi, Y. G. Yao, A. Torroni, and A. Salas. 2007. 'Distorted' mitochondrial DNA sequences in schizophrenic patients. European Journal of Human Genetics 15:400-402; author reply 402-404.

Bandelt, H. J., L. Quintana-Murci, A. Salas, and V. Macaulay. 2002. The fingerprint of phantom mutations in mitochondrial DNA data. American Journal of Human Genetics 71:1150-1160.

Bandelt, H. J., A. Salas, and C. Bravi. 2004a. Problems in FBI mtDNA database. Science 305:1402-1404.

Bandelt, H. J., A. Salas, and S. Lutz-Bonengel. 2004b. Artificial recombination in forensic mtDNA population databases. International Journal of Legal Medicine 118:267-273.

Behar, D. M., E. Metspalu, T. Kivisild, A. Achilli, Y. Hadid, S. Tzur, L. Pereira, A. Amorim, L. Quintana-Murci, K. Majamaa, C. Herrnstadt, N. Howell, O. Balanovsky, I. Kutuev, A. Pshenichnov, D. Gurwitz, B. Bonne-Tamir, A. Torroni, R. Villems, and K. Skorecki. 2006. The matrilineal ancestry of Ashkenazi Jewry: portrait of a recent founder event. American Journal of Human Genetics 78:487-497.

Bellwood, P. 1978. Man's Conquest of the Pacific - the prehistory of Southeast Asia and Oceania. Auckland: Collins.

Bendall, K. E., and B. C. Sykes. 1995. Length heteroplasmy in the first hypervariable segment of the human mtDNA control region. American Journal of Human Genetics 57:248-256.

Bensasson, D., D. Zhang, D. L. Hartl, and G. M. Hewitt. 2001. Mitochondrial pseudogenes: evolution's misplaced witnesses. Trends in Ecology and Evolution 16:314-321.

Bruen, T. C., H. Philippe, and D. Bryant. 2006. A simple and robust statistical test for detecting the presence of recombination. Genetics 172:2665-2681.

Burley, D. V., and W. R. Dickinson. 2001. Origin and significance of a founding settlement in Polynesia. Proceedings of the National Academy of Sciences U S A 98:11829-11831.

Cann, R. L., and J. K. Lum. 2004. Dispersal ghosts in Oceania. American Journal of Human Biology 16:440-451.

Cao, L., H. Shitara, T. Horii, Y. Nagao, H. Imai, K. Abe, T. Hara, J. Hayashi, and H. Yonekawa. 2007. The mitochondrial bottleneck occurs without reduction of mtDNA content in female mouse germ cells. Nature Genetics 39:386-390.

Capelli, C., J. F. Wilson, M. Richards, M. P. Stumpf, F. Gratrix, S. Oppenheimer, P. Underhill, V. L. Pascali, T. M. Ko, and D. B. Goldstein. 2001. A predominantly indigenous paternal heritage for the

Austronesian-speaking peoples of insular Southeast Asia and Oceania. American Journal of Human Genetics 68:432-443.

Clarke, A. C., M. K. Burtenshaw, P. A. McLenachan, D. L. Erickson, and D. Penny. 2006. Reconstructing the origins and dispersal of the Polynesian bottle gourd (*Lagenaria siceraria*). Molecular Biology and Evolution 23:893-900.

Clayton, D. A. 1982. Replication of animal mtDNA. Cell 28:693-705.

Clement, M., D. Posada, and K. A. Crandall. 2000. TCS: a computer program to estimate gene genealogies. Molecular Ecology 9:1657-1659.

Coble, M. D., R. S. Just, J. E. O'Callaghan, I. H. Letmanyi, C. T. Peterson, J. A. Irwin, and T. J. Parsons. 2004. Single nucleotide polymorphisms over the entire mtDNA genome that increase the power of forensic testing in Caucasians. International Journal of Legal Medicine 118:137-146.

Coskun, P. E., M. F. Beal, and D. C. Wallace. 2004. Alzheimer's brains harbor somatic mtDNA control-region mutations taht suppress mitochondrial transcription and replication. Proccedings of the National Academy of Sciences USA 101:10726-10731.

Cox, M. P. 2003. Genetic patterning at Austronesian contact zones. PhD thesis, University of Otago.

Cox, M. P. 2005. Indonesian mitochondrial DNA and its opposition to a Pleistocene era origin of proto-Polynesians in Island Southeast Asia. Human Biology 77:179-188.

Cox, M. P., and M. Mirazon Lahr. 2006. Y-chromosome diversity is inversely associated with language affiliation in paired Austronesian- and Papuan-speaking communities from Solomon Islands. American Journal of Human Biology 18:35-50.

D'Aurelio, M., C. D. Gajewski, M. T. Lin, W. M. Mauck, L. Z. Shao, G. Lenaz, C. T. Moraes, and G. Manfredi. 2004. Heterologous mitochondrial DNA recombination in human cells. Human Molecular Genetics 13:3171-3179.

Diamond, J. M. 2000. Taiwan's gift to the world. Nature 403:709-710.

Drummond, A. J., S. F. W. Ho, M. J. Phillips, and A. Rambaut. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biology 4:699-710.

Dowton, M., and N. J. Campbell. 2001. Intramitochondrial recombination - is it why some mitochondrial genes sleep around? Trends in Ecology and Evolution 16:269-271.

Dunn, M., A. Terrill, G. Reesink, R. A. Foley, and S. C. Levinson. 2005. Structural phylogenetics and the reconstruction of ancient language history. Science 309:2072-2075.

Elson, J. L., R. M. Andrews, P. F. Chinnery, R. N. Lightowlers, D. M. Turnbull and N. Howell. 2001. Analysis of European mtDNAs for recombination. American Journal of Human Genetics 68:145-153.

Erickson, D. L., B. D. Smith, A. C. Clarke, D. H. Sandweiss, and N. Tuross. 2005. An Asian origin for a 10,000-year-old domesticated plant in the Americas. Proceedings of the National Academy of Sciences of the United States of America 102:18315-18320.

Excoffier, L., G. Laval, and S. Schneider. 2005. Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online 1:47-50.

Eyre-Walker, A. 2000. Do mitochondria recombine in humans? Philosophical Transactions of the Royal

Society of London Series B: Biological Sciences355:1573-80.

Eyre-Walker, A., N. H. Smith, and J. M. Smith. 1999a. How clonal are human mitochondria? Proceedings. Biological sciences / The Royal Society 266:477-483.

Eyre-Walker, A., N. H. Smith, and J. M. Smith. 1999b. Reply to Macaulay et al. (1999): mitochondrial DNA recombination - reasons to panic. Proceedings of the Royal Society of London Series B-Biological Sciences 266:2041-2042.

Fernandez-Silva, P., J. A. Enriquez, and J. Montoya. 2003. Replication and transcription of mammalian mitochondrial DNA. Experimental Physiology 88:41-56.

Finnila, S., M. S. Lehtonen, and K. Majamaa. 2001. Phylogenetic network for European mtDNA. American Journal of Human Genetics 68:1475-1484.

Fish, J., N. Raule, and G. Attardi. 2004. Discovery of a major D-loop replication origin reveals two modes of human mtDNA synthesis. Science 306:2098-2101.

Forster, L., P. Forster, S. Lutz-Bonengel, H. Willkomm, and B. Brinkmann. 2002. Natural radioactivity and human mitochondrial DNA mutations. Proceedings of the National Academy of Sciences of the United States of America 99:13950-13954.

Forster, P. 2003. To err is human. Annals of Human Genetics 67:2-4.

Forster, P., R. Harding, A. Torroni, and H. J. Bandelt. 1996. Origin and evolution of Native American mtDNA variation: a reappraisal. American Journal of Human Genetics 59:935-45.

Friedlaender, J., T. Schurr, F. Gentz, G. Koki, F. Friedlaender, G. Horvat, P. Babb, S. Cerchio, F. Kaestle, M. Schanfield, R. Deka, R. Yanagihara, and D. A. Merriwether. 2005. Expanding southwest pacific mitochondrial haplogroups P and Q. Molecular Biology and Evolution 22:1506-17.

Friedlaender, J. S., F. R. Friedlaender, J. A. Hodgson, M. Stoltz, G. Koki, G. Horvat, S. Zhadanov, T. G. Schurr, and D. A. Merriwether. 2007. Melanesian mtDNA Complexity. PLoS ONE 2:e248.

Fris, B. N. 2006. Comparison of the efficacy of Y-SNPs, Y-STRs and mitochondrial DNA. MSc thesis, University of Auckland.

Garrido, N., L. Griparic, E. Jokitalo, J. Wartiovaara, A. M. van der Bliek, and J. N. Spelbrink. 2003. Composition and dynamics of human mitochondrial nucleoids. Molecular Biology of the Cell 14:1583-1596.

Gemmell, N. J., V. J. Metcalf and F. W. Allendorf. 2004. Mother's curse: the effect of mtDNA on individual fitness and population viability. Trends in Ecology and Evolution 19:238-244.

Gemmell, N. J., P. S. Western, J. M. Watson and J. A. M. Graves. 1996. Evolution of the mammalian mitochondrial control region - Comparisons of control region sequences between monotreme and therian mammals. Molecular Biology and Evolution 13:798-808.

Gonder, M. K., H. M. Mortensen, F. A. Reed, A. de Sousa, and S. A. Tishkoff. 2007. Whole-mtDNA genome sequence analysis of ancient African lineages. Molecular Biology and Evolution 24:757-768.

Gosden, C. 1992. Production systems and the colonization of the Western Pacific. World Archaeology 24:55-69.

Gray, R. 2005. Evolution. Pushing the time barrier in the quest for language roots. Science 309:2007-2008.

Gray, R. D., and F. M. Jordan. 2000. Language trees support the express-train sequence of Austronesian expansion. Nature 405:1052-1055.

Green, R. C. 1989. Lapita People: an introductory context for skeletal materials associated with pottery of this cultural complex. Records of the Australian Museum 41:207-213.

Green, R. C. 1991. “Near and Remote Oceania - disestablishing “Melanesia” in culture history,” in Man and a half: essays in Pacific anthropology and ethnobiology in honour of Ralph Bulmer, vol. 48. Edited by A. Pawley. Auckland: The Polynesian Society.

Green, R. C. 2003. “The Lapita horizon and traditions - Signature for one set of oceanic migrations,” in Pacific Archaeology: Assessments and Anniversary of the First Lapita Excavation (July 1952), vol. 15. Edited by C. Sand, pp. 95-120. Noumea: Le Cahiers de l’Archeologie en Nouvelle-Caledonie.

Hagelberg, E. 2003. Recombination or mutation rate heterogeneity? Implications for Mitochondrial Eve. Trends in Genetics 19:84-90.

Hagelberg, E., and J. B. Clegg. 1993. Genetic polymorphisms in prehistoric Pacific islanders determined by analysis of ancient bone DNA. Proceedings. Biological sciences / The Royal Society 252:163-170.

Hagelberg, E., N. Goldman, P. Lio, S. Whelan, W. Schiefenhovel, J. B. Clegg, and D. K. Bowden. 1999a. Evidence for mitochondrial DNA recombination in a human population of island Melanesia. Proceedings of the Royal Society of London Series B: Biological Sciences 266:485-492.

Hagelberg, E., N. Goldman, P. Lio, S. Whelan, W. Schiefenhovel, J. B. Clegg, and D. K. Bowden. 2000. Evidence for mitochondrial DNA recombination in a human population of island Melanesia: correction. Proceedings of the Royal Society of London Series B: Biological Sciences 267:1595-1596.

Hagelberg, E., M. Kayser, M. Nagy, L. Roewer, H. Zimdahl, M. Krawczak, P. Lio, and W. Schiefenhovel. 1999b. Molecular genetic evidence for the human settlement of the Pacific: analysis of mitochondrial DNA, Y chromosome and HLA markers. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences354:141-152.

Hagelberg, E., S. Quevedo, D. Turbon, and J. B. Clegg. 1994. DNA from ancient Easter Islanders. Nature 369:25-26.

Helgason, A., G. Palsson, H. S. Pedersen, E. Angulalik, E. D. Gunnarsdottir, B. Yngvadottir, and K. Stefansson. 2006. mtDNA variation in Inuit populations of Greenland and Canada: migration history and population structure. American Journal of Physical Anthropology 130:123-134.

Herrnstadt, C., J. L. Elson, E. Fahy, G. Preston, D. M. Turnbull, C. Anderson, S. S. Ghosh, J. M. Olefsky, M. F. Beal, R. E. Davis, and N. Howell. 2002. Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences for the major African, Asian, and European haplogroups. American Journal of Human Genetics 70:1152-1171.

Hertzberg, M., K. N. P. Mickleson, S. W. Serjeantson, J. F. Prior and R. J. Trent. 1989. An Asian-specific 9-bp deletion of mitochondrial DNA is frequently found in Polynesians. American Journal of Human Genetics 44: 504-510.

Hey, J. 2000. Human mitochondrial DNA recombination: can it be true? Trends in Ecology and Evolution 15:181-182.

Hill, C., P. Soares, M. Mormina, V. Macaulay, D. Clarke, P. B. Blumbach, M. Vizuete-Forster, P. Forster, D. Bulbeck, S. Oppenheimer, and M. Richards. 2007. A mitochondrial stratigraphy for island southeast

Asia. American Journal of Human Genetics 80:29-43.

Holdaway, R. N. 1996. Arrival of rats in New Zealand. Nature 384:225-226.

Holland, B., and V. Moulton. 2003. Consensus networks: A method for visualising incompatibilities in collections of trees. Algorithms in Bioinformatics, Proceedings 2812:165-176.

Holland, B. R., F. Delsuc, and V. Moulton. 2005a. Visualizing conflicting evolutionary hypotheses in large collections of trees: Using consensus networks to study the origins of placentals and hexapods. Systematic Biology 54:66-76.

Holland, B. R., K. T. Huber, D. Penny, and V. Moulton. 2005b. The MinMax squeeze: Guaranteeing a minimal tree for population data. Molecular Biology and Evolution 22:235-242.

Horai, S., K. Hayasaka, R. Kondo, K. Tsugane, and N. Takahata. 1995. Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. Proceedings of the National Academy of Sciences U S A 92:532-536.

Horai, S., K. Murayama, K. Hayasaka, S. Matsubayashi, Y. Hattori, G. Fucharoen, S. Harihara, K. S. Park, K. Omoto, and I. H. Pan. 1996. mtDNA polymorphism in East Asian Populations, with special reference to the peopling of Japan. American Journal of Human Genetics 59:579-590.

Howe, K. R. 1999. Maori/Polynesian origins and the “New Learning”. Journal of the Polynesian Society 108:305-325.

Howe, K. R.. 2003. The quest for origins. Auckland: Penguin Books (NZ) Ltd.

Howell, N., C. B. Smejkal, D. A. Mackey, P. F. Chinnery, D. M. Turnbull, and C. Herrnstadt. 2003. The pedigree rate of sequnce divergence in the human mitochondrial genome: there is a difference between phyogenetic and pedigree rates. American Journal of Human Genetics 69:1113-1126.

Howell, N., J. L. Elson, D. M. Turnbull, and C. Herrnstadt. 2004. African Haplogroup L mtDNA sequences show violations of clock-like evolution. Molecular Biology and Evolution 21:1843-1854.

Howell, N., I. Kubacka, S. M. Keers, D. M. Turnbull, and P. F. Chinnery. 2005. Co-segregation and heteroplasmy of two coding-region mtDNA mutations within a matrilineal pedigree. Human Genetics 116:28-32.

Howells, W.W. 1973. The Pacific Islanders. Wellington: Reed.

Huber, K. T., M. Langton, D. Penny, V. Moulton, and M. Hendy. 2002. Spectronet: a package for computing spectra and median networks. Applied Bioinformatics 1:159-161.

Hurles, M. E., C. Irven, J. Nicholson, P. G. Taylor, F. R. Santos, J. Loughlin, M. A. Jobling, and B. C. Sykes. 1998. European Y-chromosomal lineages in Polynesians: a contrast to the population structure revealed by mtDNA. American Journal of Human Genetics 63:1793-1806.

Hurles, M. E., E. Maund, J. Nicholson, E. Bosch, C. Renfrew, B. C. Sykes, and M. A. Jobling. 2003. Native American Y chromosomes in Polynesia: the genetic impact of the Polynesian slave trade. American Journal of Human Genetics 72:1282-1287.

Hurles, M. E., J. Nicholson, E. Bosch, C. Renfrew, B. C. Sykes, and M. A. Jobling. 2002. Y chromosomal evidence for the origins of Oceanic-speaking peoples. Genetics 160:289-303.

Hurles, M. E., B. C. Sykes, M. A. Jobling, and P. Forster. 2005. The dual origin of the Malagasy in Island Southeast Asia and East Africa: Evidence from maternal and paternal lineages. American Journal of Human Genetics 76:894-901.

Huson, D. H., and D. Bryant. 2006. Application of phylogenetic networks in evolutionary studies. Molecular Biology and Evolution 23:254-267.

Iborra, F. J., H. Kimura, and P. R. Cook. 2004. The functional organization of mitochondrial genomes in human cells. BMC Biology 2:9.

Ingman, M., and U. Gyllensten. 2003. Mitochondrial genome variation and evolutionary history of Australian and New Guinean aborigines. Genome Research 13:1600-1606.

Ingman, M., and U. Gyllensten. 2006. mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences. Nucleic Acids Research 34:D749-751.

Ingman, M., and U. Gyllensten. 2007. Rate variation between mitochondrial domains and adaptive evolution in humans. Human Molecular Genetics 16:2281-2287.

Ingman, M., H. Kaessmann, S. Paabo, and U. Gyllensten. 2000. Mitochondrial genome variation and the origin of modern humans. Nature 408:708-713.

Innan, H., and M. Nordborg. 2002. Recombination or mutational hot spots in human mtDNA? Molecular Biology and Evolution 19:1122-1127.

Jansen, R. P., and K. de Boer. 1998. The bottleneck: mitochondrial imperatives in oogenesis and ovarian follicular fate. Molecular and Cellular Endocrinology145:81-88.

Jansen, R. P. S., and G. J. Burton. 2004. Mitochondrial dysfunction in reproduction. Mitochondrion 4:577-600.

Jobling, M. A., and C. Tyler-Smith. 2003. The human Y chromosome: an evolutionary marker comes of age. Nature Reviews Genetics 4:598-612.

Kato, T., H. Kunugi, S. Nanko, and N. Kato. 2001. Mitochondrial DNA polymorphisms in bipolar disorder. J Affect Disord 62:151-164.

Kayser, M., S. Brauer, R. Cordaux, A. Casto, O. Lao, L. A. Zhivotovsky, C. Moyse-Faurie, R. B. Rutledge, W. Schiefenhoevel, D. Gil, A. A. Lin, P. A. Underhill, P. J. Oefner, R. J. Trent, and M. Stoneking. 2006. Melanesian and Asian origins of Polynesians: mtDNA and Y chromosome gradients across the Pacific. Molecular Biology and Evolution 23:2234-2244.

Kayser, M., S. Brauer, G. Weiss, W. Schiefenhovel, P. A. Underhill, and M. Stoneking. 2001. Independent histories of human Y chromosomes from Melanesia and Australia. American Journal of Human Genetics 68:173-190.

Kayser, M., S. Brauer, G. Weiss, P. A. Underhill, L. Roewer, W. Schiefenhovel, and M. Stoneking. 2000. Melanesian origin of Polynesian Y chromosomes. Current Biology 10:1237-1246.

Kazuno A. A., K. Munakata, T. Nagai, S. Shimozono, M. Tanaka, M. Yoneda, N. Kato, A. Miyawaki, and T. Kato. 2006. Identification of mitochondrial DNA polymorphisms that alter mitochondrial matrix pH and intracellular calcium dynamics. PLoS Genetics 2:1167-1177.

Ke, Y., B. Su, X. Song, D. Lu, L. Chen, H. Li, C. Qi, S. Marzuki, R. Deka, P. Underhill, C. Xiao, M. Shriver, J. Lell, D. Wallace, R. S. Wells, M. Seielstad, P. Oefner, D. Zhu, J. Jin, W. Huang, R. Chakraborty, Z. Chen, and L. Jin. 2001. African origin of modern humans in East Asia: a tale of 12,000 Y chromosomes. Science 292:1151-1153.

Kimura, M. 1968. Evolutionary rate at the molecular level. Nature 217:624-626.

Kirch, P. V. 2000. On the Road of the Winds: An Archaeological History of the Pacific Islands Before European Contact. Berkeley: University of California Press.

Kirch, P. V., and R. C. Green. 1987. History, phylogeny and evolution in Polynesia. Current Anthropology 28:431-456.

Kirch, P. V., and R. C. Green. 2001. Hawaiki, Ancestral Polynesia. Cambridge: Cambridge University Press.

Kivisild, T., P. Shen, D. P. Wall, B. Do, R. Sung, K. Davis, G. Passarino, P. A. Underhill, C. Scharfe, A. Torroni, R. Scozzari, D. Modiano, A. Coppa, P. de Knijff, M. Feldman, L. L. Cavalli-Sforza, and P. J. Oefner. 2006. The role of selection in the evolution of human mitochondrial genomes. Genetics 172:373-387.

Kong, Q. P., H. J. Bandelt, C. Sun, Y. G. Yao, A. Salas, A. Achilli, C. Y. Wang, L. Zhong, C. L. Zhu, S. F. Wu, A. Torroni, and Y. P. Zhang. 2006. Updating the East Asian mtDNA phylogeny: a prerequisite for the identification of pathogenic mutations. Hum Mol Genet 15:2076-2086.

Kong, Q. P., Y. G. Yao, M. Liu, S. P. Shen, C. Chen, C. L. Zhu, M. G. Palanichamy, and Y. P. Zhang. 2003. Mitochondrial DNA sequence polymorphisms of five ethnic populations from northern China. Human Genetics 113:391-405.

Krakauer, D. C., and A. Mira. 1999. Mitochondria and germ-cell death. Nature 400:125-126.

Kraytsberg, Y., M. Schwartz, T. A. Brown, K. Ebralidse, W. S. Kunz, D. A. Clayton, J. Vissing, and K. Khrapko. 2004. Recombination of human mitochondrial DNA. Science 304:981.

Kreitman, M. 2000. Methods to detect selection in populations with applications to the human. Annual Review of Genomics Human Genetics 1:539-559.

Kumar, S., K. Tamura, and M. Kei. 2004. MEGA 3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. Briefings in Bioinformatics 5:150-163.

Larsen, N. B., M. Rasmussen, and L. J. Rasmussen. 2005. Nuclear and mitochondrial DNA repair: similar pathways? Mitochondrion 5:89-108.

Larson, G., T. Cucchi, M. Fujita, E. Matisoo-Smith, J. Robins, A. Anderson, B. Rolett, M. Spriggs, G. Dolman, T. H. Kim, N. T. Thuy, E. Randi, M. Doherty, R. A. Due, R. Bollt, T. Djubiantono, B. Griffin, M. Intoh, E. Keane, P. Kirch, K. T. Li, M. Morwood, L. M. Pedrina, P. J. Piper, R. J. Rabett, P. Shooter, G. Van den Bergh, E. West, S. Wickler, J. Yuan, A. Cooper, and K. Dobney. 2007. Phylogeny and ancient DNA of Sus provides insights into neolithic expansion in Island Southeast Asia and Oceania. Proceedings of the National Academy of Sciences U S A 104:4834-4839.

Legros, F., F. Malka, P. Frachon, A. Lombes, and M. Rojo. 2004. Organization and dynamics of human mitochondrial DNA. Journal of Cell Science 117:2653-2662.

Lewis, C. M., Jr., R. Y. Tito, B. Lizarraga, and A. C. Stone. 2005. Land, language, and loci: mtDNA in Native Americans and the genetic history of Peru. American Journal of Physical Anthropology 127:351-360.

Lie, B. A., B. M. Dupuy, A. Spurkland, M. A. Fernandez-Vina, E. Hagelberg and E. Thorsby. 2007. Molecular genetic studies of an early European and Amerindian contribution to the Polynesian gene pool. Tissue Antigens 69:10-18.

Lightowlers, R. N., P. F. Chinnery, D. M. Turnbull, and N. Howell. 1997. Mammalian mitochondrial genetics: heredity, heteroplasmy and disease. Trends in Genetics 13:450-455.

Liston, J. 2005. An assessment of radiocarbon dates from Palau, Western Micronesia. Radiocarbon 47:295-354.

Lum, J. K., and R. L. Cann. 1998. mtDNA and language support a common origin of Micronesians and Polynesians in Island Southeast Asia. American Journal of Physical Anthropology 105:109-119.

Lum, J. K., R. L. Cann, J. J. Martinson, and L. B. Jorde. 1998. Mitochondrial and nuclear genetic relationships among Pacific Island and Asian populations. American Journal of Human Genetics 63:613-624.

Lum, J. K., O. Rickards, C. Ching, and R. L. Cann. 1994. Polynesian mitochondrial DNAs reveal three deep maternal lineage clusters. Human Biology 66:567-590.

Maca-Meyer, N., A. M. Gonzalez, J. M. Larruga, C. Flores, and V. M. Cabrera. 2001. Major genomic mitochondrial lineages delineate early human expansions. BMC Genetics 2:13.

Maca-Meyer, N., A. M. Gonzalez, J. Pestano, C. Flores, J. M. Larruga, and V. M. Cabrera. 2003. Mitochondrial DNA transit between West Asia and North Africa inferred from U6 phylogeography. BMC Genet 4:15.

Macaulay, V., C. Hill, A. Achilli, C. Rengo, D. Clarke, W. Meehan, J. Blackburn, O. Semino, R. Scozzari, F. Cruciani, A. Taha, N. K. Shaari, J. M. Raja, P. Ismail, F. Zainuddin, W. Goodwin, D. Bulbeck, H. J. Bandelt, S. Oppenheimer, A. Torroni, and M. Richards. 2005a. Tracing modern human origins - Response. Science 309:1995-1996.

Macaulay, V., C. Hill, A. Achilli, C. Rengo, D. Clarke, W. Meehan, J. Blackburn, O. Semino, R. Scozzari, F. Cruciani, A. Taha, N. K. Shaari, J. M. Raja, P. Ismail, Z. Zainuddin, W. Goodwin, D. Bulbeck, H. J. Bandelt, S. Oppenheimer, A. Torroni, and M. Richards. 2005b. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science 308:1034-1036.

Macaulay, V., M. Richards, and B. Sykes. 1999. Mitochondrial DNA recombination-no need to panic. Proceedings of the Royal Society of London Series B: Biological Sciences 266:2037-9; discussion 2041-2042.

Malyarchuk, B. A. 2004. Similarity of mutation spectra of the mitochondrial DNA hypervariable segment 1 in Homo and Pan species. Molecular Biology 38:370-375.

Malyarchuk, B. A., I. B. Rogozin, V. B. Berikov, and M. V. Derenko. 2002. Analysis of phylogenetically reconstructed mutational spectra in human mitochondrial DNA control region. Human Genetics 111:46-53.

Martinson, J. J., R. M. Harding, G. Philippon, F. F. Sainte-Marie, J. Roux, A. J. Boyce, and J. B. Clegg. 1993. Demographic reductions and genetic bottlenecks in humans: minisatellite allele distributions in Oceania. Human Genetics 91:445-450.

Matisoo-Smith, E. 2002. Something old, something new: do genetic studies of contemporary populations reliably represent prehistoric populations of Pacific Rattus exulans? Human Biology 74:489-496.

Matisoo-Smith, E., R. M. Roberts, G. J. Irwin, J. S. Allen, D. Penny, and D. M. Lambert. 1998. Patterns of prehistoric human mobility in Polynesia indicated by mtDNA from the Pacific rat. Proceedings of the National Academy of Sciences U S A 95:15145-15150.

Matisoo-Smith, E., and J. H. Robins. 2004. Origins and dispersals of Pacific peoples: evidence from mtDNA phylogenies of the Pacific rat. Proceedings of the National Academy of Sciences U S A 101:9167-9172.

May-Panloup, P., M. F. Chretien, Y. Malthiery, and P. Reynier. 2007. Mitochondrial DNA in the oocyte and the developing embryo. Current Topics in Developmental Biology 77:51-83.

McDonald, J. H., and M. Kreitman. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652-654.

McFarland R., J. L. Elson, R. W. Taylor, N. Howell, and D. M. Turnbull. 2004. Assigning pathogenicity to mitochondrial tRNA mutations: when 'definitely maybe' is not good enough. Trends in Genetics 20:591-596.

Melton, P. E., I. Briceno, A. Gomez, E. J. Devor, J. E. Bernal, and M. H. Crawford. 2007. Biological relationship between Central and South American Chibchan speaking populations: evidence from mtDNA. American Journal of Physical Anthropology 133:753-770.

Melton, T., R. Peterson, A. J. Redd, N. Saha, A. S. Sofro, J. Martinson, and M. Stoneking. 1995. Polynesian genetic affinities with Southeast Asian populations as identified by mtDNA analysis. American Journal of Human Genetics 57:403-414.

Merriwether, D. A., J. A. Hodgson, F. R. Friedlaender, R. Allaby, S. Cerchio, G. Koki, and J. S. Friedlaender. 2005. Ancient mitochondrial M haplogroups identified in the Southwest Pacific. Proceedings of the National Academy of Sciences U S A 102:13034-13039.

Meyer, S., G. Weiss, and A. von Haeseler. 1999. Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA. Genetics 152:1103-1110.

Mishmar, D., E. Ruiz-Pesini, P. Golik, V. Macaulay, A. G. Clark, S. Hosseini, M. Brandon, K. Easley, E. Chen, M. D. Brown, R. I. Sukernik, A. Olckers, and D. C. Wallace. 2003. Natural selection shaped regional mtDNA variation in humans. Proceedings of the National Academy of Sciences U S A 100:171-176.

Moilanen, J. S., and K. Majamaa. 2003. Phylogenetic network and physicochemical properties of nonsynonymous mutations in the protein-coding genes of human mitochondrial DNA. Molecular Biology and Evolution 20:1195-1210.

Murray-McIntosh, R. P., B. J. Scrimshaw, P. J. Hatfield, and D. Penny. 1998. Testing migration patterns and estimating founding population size in Polynesia by using human mtDNA sequences. Proceedings of the National Academy of Sciences U S A 95:9047-9052.

Niemi, A. K., J. S. Moilanen, M. Tanaka, A. Hervonen, M. Hurme, T. Lehtimaki, Y. Arai, N. Hirose, and K. Majamaa. 2005. A combination of three common inherited mitochondrial DNA polymorphisms promotes longevity in Finnish and Japanese subjects. European Journal of Human Genetics 13:166-170.

Nishimura, Y., T. Yoshinari, K. Naruse, T. Yamada, K. Sumi, H. Mitani, T. Higashiyama, and T. Kuroiwa. 2006. Active digestion of sperm mitochondrial DNA in single living sperm revealed by optical tweezers. Proceedings of the National Academy of Sciences U S A 103:1382-1387.

O'Connell, J. F., and J. Allen. 2004. Dating the colonization of Sahul (Pleistocene Australia-New Guinea): a review of recent research. Journal of Archaeological Science 31:835-853.

Ohashi, J., I. Naka, K. Tokunaga, T. Inaoka, Y. Ataka, M. Nakazawa, Y. Matsumura, and R. Ohtsuka. 2006. Brief communication: Mitochondrial DNA variation suggests extensive gene flow from Polynesian ancestors to indigenous Melanesians in the northwestern Bismarck Archipelago. American Journal of Physical Anthropology 130:551-556.

Oota, H., T. Kitano, F. Jin, I. Yuasa, L. Wang, S. Ueda, N. Saitou, and M. Stoneking. 2002. Extreme mtDNA homogeneity in continental Asian populations. American Journal of Physical Anthropology 118:146-153.

Oppenheimer, S. 2004. The 'Express Train from Taiwan to Polynesia': on the congruence of proxy lines of evidence. World Archaeology 36:591-600.

Palanichamy, M. G., C. Sun, S. Agrawal, H. J. Bandelt, Q. P. Kong, F. Khan, C. Y. Wang, T. K. Chaudhuri, V. Palla, and Y. P. Zhang. 2004. Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing: Implications for the peopling of South Asia. American Journal of Human Genetics 75:966-978.

Parr, R. L., J. Maki, B. Reguly, G. D. Dakubo, A. Aguirre, R. Wittock, K. Robinson, J. P. Jakupciak, and R. E. Thayer. 2006. The pseudo-mitochondrial genome influences mistakes in heteroplasmy interpretation. BMC Genomics 7:185.

Pawley, A., and R. C. Green. 1973. Dating the dispersal of the Oceanic languages. Oceanic Linguistics 12:1-67.

Pereira, L., J. Goncalves, R. Franco-Duarte, J. Silva, T. Rocha, C. Arnold C , M. Richards and V. Macaulay. 2007. No evidence for an mtDNA role in sperm motility: Data from complete sequencing of asthenozoospermic males. Molecular Biology and Evolution 24:868-874.

Pierson, M. J., R. Martinez-Arias, B. R. Holland, N. J. Gemmell, M. E. Hurles, and D. Penny. 2006. Deciphering past human population movements in Oceania: Provably optimal trees of 127 mtDNA genomes. Molecular Biology and Evolution 23:1966-1975.

Piganeau, G., and A. Eyre-Walker. 2004. A reanalysis of the indirect evidence for recombination in human mitochondrial DNA. Heredity 92:282-288.

Quintans, B., V. Alvarez-Iglesias, A. Salas, C. Phillips, M. V. Lareu, and A. Carracedo. 2004. Typing of mitochondrial DNA coding region SNPs of forensic and anthropological interest using SNaPshot minisequencing. Forensic Science International 140:251-257.

Fischer, S.R. 2002. A history of the Pacific Islanders. New York: Palgrave.

Rajkumar, R., J. Banerjee, H. B. Gunturi, R. Trivedi, and V. K. Kashyap. 2005. Phylogeny and antiquity of M macrohaplogroup inferred from complete mt DNA sequence of Indian specific lineages. BMC Evolutionary Biology 5:26.

Rambaut, A. 1996. "Se-Al: Sequence Alignment Editor, v2.0a11," Available at: http://evolve.zoo.ox.ac.uk/.

Redd, A. J., N. Takezaki, S. T. Sherry, S. T. McGarvey, A. S. Sofro, and M. Stoneking. 1995. Evolutionary history of the COII/tRNALys intergenic 9 base pair deletion in human mitochondrial DNAs from the Pacific. Molecular Biology and Evolution 12:604-615.

Reynier, P., P. May-Panloup, M. F. Chretien, C. J. Morgan, M. Jean, F. Savagner, P. Barriere, and Y. Malthiery. 2001. Mitochondrial DNA content affects the fertilizability of human oocytes. Molecular Human Reproduction 7:425-429.

Richards, M., S. Oppenheimer, and B. Sykes. 1998. mtDNA suggests Polynesian origins in Eastern Indonesia. American Journal of Human Genetics 63:1234-1236.

Rieder, M. J., S. L. Taylor, V. O. Tobe, and D. A. Nickerson. 1998. Automating the identification of DNA

variations using quality-based fluorescence re-sequencing: analysis of the human mitochondrial genome. Nucleic Acids Research 26:967-973.

Roostalu, U., I. Kutuev, E. L. Loogvali, E. Metspalu, K. Tambets, M. Reidla, E. K. Khusnutdinova, E. Usanga, T. Kivisild, and R. Villems. 2007. Origin and expansion of haplogroup H, the dominant human mitochondrial DNA lineage in West Eurasia: the Near Eastern and Caucasian perspective. Molecular Biology and Evolution 24:436-448.

Rowold, D. J., J. R. Luis, M. C. Terreros, and R. J. Herrera. 2007. Mitochondrial DNA gene flow indicates preferred usage of the Levant Corridor over the Horn of Africa passageway. Journal of Human Genetics 52:436-447.

Rozas, J., J. C. Sanchez-DelBarrio, X. Messeguer, and R. Rozas. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496-2497.

Ruiz-Pesini, E., D. Mishmar, Brandon M., Procaccio V., and D. C. Wallace. 2004. Effects of purifying and adaptive selection on regional variation in human mtDNA. Science 303:223-226.

Ruiz-Pesini, E., A. C. Lapena, C. Diez-Sanchez, A. Perez-Martos, J. Montoya, E. Alvarez, M. Diaz, A. Urrieis, L. Montoro, M. J. Lopez-Perez and J. A. Enriquez. 2000. Human mtDNA haplogroups associated with high or reduced spermatozoa motility. American Journal of Human Genetics 67:682-696.

Saillard, J., P. Forster, N. Lynnerup, H. J. Bandelt, and S. Norby. 2000. mtDNA variation among Greenland Eskimos: The edge of the Beringian expansion. American Journal of Human Genetics 67:718-726.

Sato, A., K. Nakada, M. Akimoto, K. Ishikawa, T. Ono, H. Shitara, H. Yonekawa, and J. L. Hayashi. 2005. Rare creation of recombinant mtDNA haplotypes in mammalian tissues. Proceedings of the National Academy of Sciences of the United States of America 102:6057-6062.

Savolainen, P., T. Leitner, A. N. Wilton, E. Matisoo-Smith, and J. Lundeberg. 2004. A detailed picture of the origin of the Australian dingo, obtained from the study of mitochondrial DNA. Proceedings of the National Academy of Sciences of the United States of America 101:12387-12390.

Saxena, R., P. I. W. de Bakker, K. Singer, V. Mootha, N. Burtt, J. N. Hirschhorn, D. Gaudet, B. Isomaa, M. J. Daly, L. Groop, K. G. Ardlie, and D. Altshuler. 2006. Comprehensive association testing of common mitochondrial DNA variation in metabolic disease. American Journal of Human Genetics 79:54-61.

Schwartz, M., and J. Vissing. 2002. Paternal inheritance of mitochondrial DNA. New England Journal of Medicine 347:576-580.

Semple, C., and M. Steel. 2003. Phylogenetics. Oxford lecture series in mathematics and its applications. New York: Oxford University Press.

Skaletsky, H., T. Kuroda-Kawaguchi, P. J. Minx, H. S. Cordum, L. Hillier, L. G. Brown, S. Repping, T. Pyntikova, J. Ali, T. Bieri, A. Chinwalla, A. Delehaunty, K. Delehaunty, H. Du, G. Fewell, L. Fulton, R. Fulton, T. Graves, S. F. Hou, P. Latrielle, S. Leonard, E. Mardis, R. Maupin, J. McPherson, T. Miner, W. Nash, C. Nguyen, P. Ozersky, K. Pepin, S. Rock, T. Rohlfing, K. Scott, B. Schultz, C. Strong, A. Tin-Wollam, S. P. Yang, R. H. Waterston, R. K. Wilson, S. Rozen, and D. C. Page. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature 423:825-837.

Slate, J., and N. J. Gemmell. 2004. Eve ‘n’ Steve: recombination of human mitochondrial DNA. Trends in

Ecology & Evolution 19:561-563.

Spurdle, A. B., D. G. Woodfield, M. F. Hammer, and T. Jenkins. 1994. The genetic affinity of Polynesians: evidence from Y chromosome polymorphisms. Annals of Human Genetics 58: 251-263.

Starikovskaya, E. B., R. I. Sukernik, O. A. Derbeneva, N. V. Volodko, E. Ruiz-Pesini, A. Torroni, M. D. Brown, M. T. Lott, S. H. Hosseini, K. Huoponen, and D. C. Wallace. 2005. Mitochondrial DNA diversity in indigenous populations of the southern extent of Siberia, and the origins of Native American haplogroups. Annals of Human Genetics 69:67-89.

Stoneking, M. 2000. Hypervariable sites in the mtDNA control region are mutational hotspots. American Journal of Human Genetics 67:1029-1032.

Storey, A. A., J. M. Ramirez, D. Quiroz, D. V. Burley, D. J. Addison, R. Walter, A. J. Anderson, T. L. Hunt, J. S. Athens, L. Huynen, and E. A. Matisoo-Smith. 2007. Radiocarbon and DNA evidence for a pre-Columbian introduction of Polynesian chickens to Chile. Proceedings of the National Academy of Sciences U S A 104:10335-10339.

Stuart, J. A., and M. F. Brown. 2006. Mitochondrial DNA maintenance and bioenergetics. Biochimica et Biophysica Acta 1757:79-89.

Su, B., L. Jin, P. Underhill, J. Martinson, N. Saha, S. T. McGarvey, M. D. Shriver, J. Chu, P. Oefner, R. Chakraborty, and R. Deka. 2000. Polynesian origins: insights from the Y chromosome. Proceedings of the National Academy of Sciences U S A 97:8225-8228.

Sun, C., Q. P. Kong, M. G. Palanichamy, S. Agrawal, H. J. Bandelt, Y. G. Yao, F. Khan, C. L. Zhu, T. K. Chaudhuri, and Y. P. Zhang. 2006. The dazzling array of basal branches in the mtDNA macrohaplogroup M from India as inferred from complete genomes. Molecular Biology and Evolution 23:683-690.

Sutovsky, P., K. Van Leyen, T. McCauley, B. N. Day, and M. Sutovsky. 2004. Degradation of paternal mitochondria after fertilization: implications for heteroplasmy, assisted reproductive technologies and mtDNA inheritance. Reproductive Biomedicine Online 8:24-33.

Swofford, D. L. 2003. “PAUP* Phylogenetic Analysis Using Parsimony (*and Other Methods),” 4th edition. Sunderland, Massachusetts: Sinauer Associates.

Sykes, B., A. Leiboff, J. Low-Beer, S. Tetzner, and M. Richards. 1995. The origins of the Polynesians: an interpretation from mitochondrial lineage analysis. American Journal of Human Genetics 57:1463-1475.

Taanman, J. W. 1999. The mitochondrial genome: structure, transcription, translation and replication. Biochimica Et Biophysica Acta-Bioenergetics 1410:103-123.

Tanaka M., V. M. Cabrera, A. M. Gonzalez, J. M. Larruga, T. Takeyasu, N. Fuku, L. J. Guo, R. Hirose, Y. Fujita, M. Kurata, K. Shinoda, K. Umetsu, Y. Yamada, Y. Oshida, Y. Sato, N. Hattori, Y. Mizuno, Y. Arai, N. Hirose, S. Ohta, O. Ogawa, Y. Tanaka, R. Kawamori, M. Shamoto-Nagai, W. Maruyama, H. Shimokata, R. Suzuki, H. Shimodaira. 2004. Mitochondrial genome variation in Eastern Asia and the peopling of Japan. Genome Research 14:1832-1850.

Tajima, A., M. Hayami, K. Tokunaga, T. Juji, M. Matsuo, S. Marzuki, K. Omoto, and S. Horai. 2004. Genetic origins of the Ainu inferred from combined DNA analyses of maternal and paternal lineages. Journal of Human Genetics 49:187-193.

Tajima, A., C. S. Sun, I. H. Pan, T. Ishida, N. Saitou, and S. Horai. 2003. Mitochondrial DNA polymorphisms in nine aboriginal groups of Taiwan: implications for the population history of aboriginal Taiwanese.

Human Genetics 113:24-33.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585-595.

Tamura, K., M. Nei, and S. Kumar. 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proceedings of the National Academy of Sciences U S A 101:11030-11035.

Taylor, R. W., and D. M. Turnbull. 2005. Mitochondrial DNA mutations in human disease. Nature Reviews Genetics 6:389-402.

Terrell, J. E. 1986. Prehistory in the Pacific Islands. Cambridge: Cambridge University Press.

Terrell, J. E. 2004. The 'sleeping giant' hypothesis and New Guinea's place in the prehistory of Greater Near Oceania. World Archaeology 36:601-609.

Thangaraj, K., G. Chaubey, T. Kivisild, A. G. Reddy, V. K. Singh, A. A. Rasalkar, and L. Singh. 2005. Reconstructing the origin of Andaman Islanders. Science 308:996.

Thyagarajan, B., R. A. Padua, and C. Campbell. 1996. Mammalian mitochondria possess homologous DNA recombination activity. Journal of Biological Chemistry 271:27536-27543.

Tommaseo-Ponzetta, M., M. Attimonelli, M. De Robertis, F. Tanzariello, and C. Saccone. 2002. Mitochondrial DNA variability of West New Guinea populations. American Journal of Physical Anthropology 117:49-67.

Torroni, A., A. Achilli, V. Macaulay, M. Richards, and H. J. Bandelt. 2006. Harvesting the fruit of the human mtDNA tree. Trends in Genetics 22:339-345.

Torroni, A., C. Rengo, V. Guida, F. Cruciani, D. Sellitto, A. Coppa, F. L. Calderon, B. Simionati, G. Valle, M. Richards, V. Macaulay, and R. Scozzari. 2001. Do the four clades of the mtDNA haplogroup L2 evolve at different rates? American Journal of Human Genetics 69:1348-1356.

Trejaut, J. A., T. Kivisild, J. H. Loo, C. L. Lee, C. L. He, C. J. Hsu, Z. Y. Li, and M. Lin. 2005. Traces of archaic mitochondrial lineages persist in Austronesian-speaking Formosan populations. PLoS Biology 3:1362-1372.

Underhill, P. A., P. Shen, A. A. Lin, L. Jin, G. Passarino, W. H. Yang, E. Kauffman, B. Bonne-Tamir, J. Bertranpetit, P. Francalacci, M. Ibrahim, T. Jenkins, J. R. Kidd, S. Q. Mehdi, M. T. Seielstad, R. S. Wells, A. Piazza, R. W. Davis, M. W. Feldman, L. L. Cavalli-Sforza, and P. J. Oefner. 2000. Y chromosome sequence variation and the history of human populations. Nature Genetics 26:358-361.

Underhill, P. A., G. Passarino, A. A. Lin, S. Marzuki, P. J. Oefner, L. L. Cavalli-Sforza and G. K. Chambers. 2001. Maori origins, Y-chromosome haplotypes and implications for human history in the Pacific. Human Mutation 17:271-280.

Van Blerkom, J. 2004. Mitochondria in human oogenesis and preimplantation embryogenesis: engines of metabolism, ionic regulation and developmental competence. Reproduction 128:269-280.

Van der Walt, J. M., K. K. Nicodemus, E. R. Martin, W. K. Scott, M. A. Nance, R. L. Watts, J. P. Hubble, J. L. Haines, W. C. Koller, K. Lyons, R. Pahwa, M. B. Stern, A. Colcher, B. C. Hiner, J. Jankovic, W. G. Ondo, F. H. Allen, Jr., C. G. Goetz, G. W. Small, F. Mastaglia, J. M. Stajich, A. C. McLaurin, L. T. Middleton, B. L. Scott, D. E. Schmechel, M. A. Pericak-Vance, and J. M. Vance. 2003. Mitochondrial polymorphisms significantly reduce the risk of Parkinson disease. American Journal of Human Genetics

72:804-811.

Van Holst Pellekaan, S. M., M. Ingman, J. Roberts-Thomson, and R. M. Harding. 2006. Mitochondrial genomics identifies major haplogroups in Aboriginal Australians. American Journal of Physical Anthropology 131:282-294.

Wen, B., H. Li, S. Gao, X. Y. Mao, Y. Gao, F. Li, F. Zhang, Y. G. He, Y. L. Dong, Y. J. Zhang, W. Huang, J. Z. Jin, C. J. Xiao, D. R. Lu, R. Chakraborty, B. Su, R. Deka, and L. Jin. 2005. Genetic structure of Hmong-Mien speaking populations in East Asia as revealed by mtDNA lineages. Molecular Biology and Evolution 22:725-734.

White, J. P., A. J., and S. J. 1988. Peopling the Pacific: the Lapita Homeland Project. Australian Natural History 22:410-416.

Whyte, A. L., S. J. Marshall, and G. K. Chambers. 2005. Human evolution in Polynesia. Human Biology 77:157-177.

Wilmshurst, J. M., and T. F. G. Higham. 2004. Using rat-gnawed seeds to independently date the arrival of Pacific rats and humans in New Zealand. Holocene 14:801-806.

Xu, X., and U. Arnason. 1996. The mitochondrial DNA molecule of Sumatran orangutan and a molecular proposal for two (Bornean and Sumatran) species of orangutan. Journal of Molecular Evolution 43:431-437.

Xu, X., and U. Arnason. 1996. A complete sequence of the mitochondrial genome of the Western lowland gorilla. Molecular Biology and Evolution 13:691-698.

Yao, Y.-G., V. Macaulay, T. Kivisild, Y.-P. Zhang, and H.-J. Bandelt. 2003. To trust or not to trust an idiosyncratic mitochondrial data set. American Journal of Human Genetics 72:1341-1346.

Zsurka, G., K. G. Hampel, T. Kudina, C. Kornblum, Y. Kraytsberg, C. E. Elger, K. Khrapko, and W. S. Kunz. 2007. Inheritance of mitochondrial DNA recombinants in double-heteroplasmic families: potential implications for phylogenetic analysis. American Journal of Human Genetics 80:298-305.

Zsurka, G., Y. Kraytsberg, T. Kudina, C. Kornblum, C. E. Elger, K. Khrapko, and W. S. Kunz. 2005. Recombination of mitochondrial DNA in skeletal muscle of individuals with multiple mitochondrial DNA heteroplasmy. Nature Genetics 37:873-877.

Deciphering Past Human Population Movements in Oceania: Provably Optimal Trees of 127 mtDNA Genomes

Melanie J. Pierson,† Rosa Martinez-Arias,‡ Barbara R. Holland,* Neil J. Gemmell,† Matthew E. Hurles,§ and David Penny**

*Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand; †School of Biological Sciences, University of Canterbury, Christchurch, New Zealand; ‡GBF German Research Centre for Biotechnology, Braunschweig, Germany; §Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, United Kingdom

The settlement of the many island groups of Remote Oceania occurred relatively late in prehistory, beginning approximately 3,000 years ago when people sailed eastwards into the Pacific from Near Oceania, where evidence of human settlement dates from as early as 40,000 years ago. Archeological and linguistic analyses have suggested the settlers of Remote Oceania had ancestry in Taiwan, as descendants of a proposed Neolithic expansion that began approximately 5,500 years ago. Other researchers have suggested that the settlers were descendants of peoples from Island Southeast Asia or the existing inhabitants of Near Oceania alone. To explore patterns of maternal descent in Oceania, we have assembled and analyzed a data set of 137 mitochondrial DNA (mtDNA) genomes from Oceania, Australia, Island Southeast Asia, and Taiwan that includes 19 sequences generated for this project. Using the MinMax Squeeze Approach (MMS), we report the consensus network of 165 most parsimonious trees for the Oceanic data set, increasing by many orders of magnitude the numbers of trees for which a provable minimal solution has been found. The new mtDNA sequences highlight the limitations of partial sequencing for assigning sequences to haplogroups and dating recent divergence events. The provably optimal trees found for the entire mtDNA sequences using the MMS method provide a reliable and robust framework for the interpretation of evolutionary relationships and confirm that the female settlers of Remote Oceania descended from both the existing inhabitants of Near Oceania and more recent migrants into the region.

Introduction

Exploring the timing and pathways of Oceanic settlement is a multidisciplinary endeavor, with the direct evidence of prehistoric populations obtained from archeological investigations increasingly supplemented by linguistic and biological studies of their present-day descendants (Hurles et al. 2003). The region of Near Oceania (fig. 1) encompasses New Guinea, the Bismarck Archipelago, Bougainville, and the northern Solomon Islands and delineates the extent of early migrations into Oceania, with the first evidence of human settlement dating from at least 44,000 years ago (O'Connell and Allen 2004). In contrast, the settlement of the many islands of Remote Oceania began just over 3,000 years ago, following the appearance in Near Oceania of an archeological horizon known as the Lapita Cultural Complex. Similar assemblages appear in the first settlement sites in Remote Oceania to the south and east from about 3,100 years ago, and a rapid settlement sequence of Remote Oceania follows, with migrants reaching the 3 points of the Polynesian triangle at Hawaii, Easter Island, and New Zealand from ~1,500 to 800 years ago (Kirch 2000).

In a recent review, Green (2003) summarizes the many models proposed for the development of the Lapita Cultural Complex in Near Oceania, describing this and the subsequent settlement of Remote Oceania as "one set of human migrations among the many that have occurred throughout Oceania during the last 40,000 and perhaps 50,000 years." Much attention has been focused on this later phase of Oceanic prehistory, and current debate centers on whether the Lapita sites represent an intrusive migration of peoples into Near Oceania and, if so, the geographic origin of these migrants and the extent of interactions between the newcomers and the existing inhabitants of Near Oceania.

All of the languages of Remote Oceania and many from Near Oceania belong to the Oceanic subgroup of the Austronesian language family. The high level of diversity among the many non-Austronesian (Papuan) languages spoken in parts of Near Oceania suggests that they have developed over a much greater time frame than the Austronesian languages, consistent with the early settlement dates in this region. Phylogenetic analysis of the Austronesian language family, widely spoken today in Oceania, Island Southeast Asia, and Taiwan has demonstrated that the language tree fits an "out-of-Taiwan" sequence of expansion (Gray and Jordan 2000). This model proposes a migration of proto–Austronesian-speaking peoples into Island Southeast Asia and Oceania from Taiwan, beginning from about 5,500 years ago (Bellwood 1991, 2001). Under this model, the Lapita sites in Near Oceania are viewed as evidence of an intrusive "Austronesian" settlement. Although early descriptions of this model presented the sequence of settlement as so rapid that it allowed very little interaction between settlers and the existing inhabitants of Near Oceania (often described as the "Express Train" model; Diamond 1988), more recent formulations emphasize integration between the 2 groups (Green 2003).

This report focuses on the inferences for Oceanic prehistory preserved in phylogenies of maternally inherited mitochondrial DNA (mtDNA), using entire genome sequences from individuals from Oceania, Australia, Island Southeast Asia, and Taiwan. Previous analyses of the hypervariable region-I (HVR-I) of the control region of mtDNA (Sykes et al. 1995; Lum et al. 1998; Murray-McIntosh et al. 1998; Hagelberg et al. 1999; Friedlaender et al. 2002) have found a general pattern of few, closely related lineages present among the populations of Remote Oceania, particularly in Polynesia, contrasting with a large number of diverse haplotypes in Near Oceanic populations.

Key words: human, mtDNA, Oceania, MMS, prehistory.

E-mail: mjp110@student.canterbury.ac.nz.

Mol. Biol. Evol. 23(10):1966–1975. 2006
doi:10.1093/molbev/msl063
Advance Access publication July 19, 2006

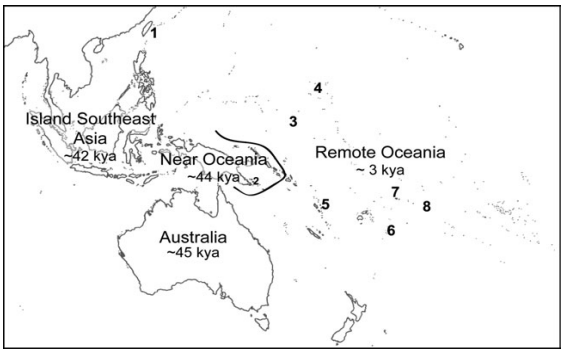


FIG. 1.—Map of Oceania. Dates provided are the earliest archeological evidence of anatomically modern human settlement (O'Connell and Allen 2004). Sources of new mtDNA sequences are numbered as follows: 1, Taiwan; 2, Trobriand Islands; 3, Kapingamarangi Atoll; 4, Majuro Atoll; 5, Vanuatu; 6, Tonga; 7, Samoa; and 8, Cook Islands.

Table 1
New Sample Details

Accession Number	Geographic Origin	Haplogroup
DQ372868	Taiwan[a]	M/M7c
DQ372869	Taiwan[a]	N/R/B5a
DQ372870	Trobriand Islands, Papua New Guinea[b]	N/R/P2
DQ372871	Trobriand Islands, Papua New Guinea[b]	N/R/B4a1a
DQ372872	Trobriand Islands, Papua New Guinea[b]	N/R/P2
DQ372873	Trobriand Islands, Papua New Guinea[b]	N/R/B4a1a1
DQ372874	Kapingamarangi Atoll, Caroline Islands[a]	N/R/B4a1a1
DQ372875	Kapingamarangi Atoll, Caroline Islands[a]	N/R/B4a1a1
DQ372876	Majuro Atoll, Marshall Islands[a]	M/M7c
DQ372877	Majuro Atoll, Marshall Islands[a]	N/R/B4a1a1
DQ372878	Espiritu Santo, Vanuatu[c]	N/R/B4a1a1/Pol. motif
DQ372879	Espiritu Santo, Vanuatu[c]	M/M28
DQ372880	Espiritu Santo, Vanuatu[c]	M/Q1
DQ372881	Maewo, Vanuatu[c]	N/R/B4a1a1/Pol. motif
DQ372882	Vanuatu[a]	M/Q1
DQ372883	Vanuatu[a]	M/M28
DQ372884	Cook Islands[c]	M/Q1
DQ372885	Samoa[a]	M/Q1
DQ372886	Tonga[a]	N/R/B4a1a1/Pol. motif
DQ372887	New Zealand (European)[d]	N/W

[a] Supplied by JB Clegg, Weatherall Institute of Molecular Medicine, University of Oxford.
[b] Provided by W Schievenhovel.
[c] Samples from M.E.H.
[d] Methodological control sample, provided by PA McLenachan.

One HVR-I haplotype has become known as the "Polynesian motif" due to its high frequency among Polynesian peoples and analyses of the distribution of this haplotype, and its immediate precursors in Oceania, Island Southeast Asia, and mainland Asia have been interpreted as supporting an out-of-Taiwan model of migration into the Pacific (Melton et al. 1995; Redd et al. 1995). However, one molecular dating estimate on this region of the mtDNA control region suggested that the time to the most recent common ancestor (TMRCA) for this motif in eastern Indonesia is ~17,000 years, beyond the ~5,500 year time frame for migration from Taiwan (Richards et al. 1998). This has lent some support to a model of origins of Lapita migrants in Island Southeast Asia, not Taiwan (Richards et al. 1998; Oppenheimer 2004); however, see Cox (2005) for a reassessment with a larger sample set and also Penny (2005) for a discussion of issues surrounding molecular dating of events in the recent past.

A recent study of entire mtDNA sequences from Taiwanese individuals with the ancestral Polynesian motif has revealed a close relationship between the Oceanic Polynesian motif haplotypes and the Taiwanese sequences (Trejaut et al. 2005). Entire mtDNA sequences from Australian and Near Oceanic populations are also beginning to reveal a number of deep lineages found only in this part of the world, which descend from the M, N, and N/R macrohaplogroups, reflecting the complexity expected of a region with >40,000 years of human prehistory (Ingman and Gyllensten 2003; Friedlaender et al. 2005; Merriwether et al. 2005).

To examine patterns of prehistoric migrations in Oceania, we have sequenced 19 mtDNA genomes from Taiwan and Oceania (table 1) and analyzed these together with over 100 other sequences from Taiwan, Island Southeast Asia, and Australia and Oceania. Steel and Penny (2004) have shown that for "ample" data (sequences that can be connected with steps of size 1) maximum parsimony and most parsimonious likelihood are equivalent. This suggests that maximum parsimony may be an appropriate criterion for analyzing densely sampled population data. The large number of unique sequences in our data set means it is not feasible to explore all possible trees and evaluate their parsimony score, and a heuristic approach is required. However, in many cases, we can guarantee that a particular tree, or set of trees, is globally optimal under the maximum parsimony criterion by using the recently described MinMax Squeeze Approach (MMS) (Holland, Huber, et al. 2005). The method takes the length of the shortest tree found by heuristic search as an upper bound and derives a lower bound by summing the parsimony scores of partitions of columns of the data set. Heuristic search is used to optimize the partition to give the highest possible lower bound. A tree or set of trees is guaranteed optimal if the upper and lower bounds meet. The MinMax Squeeze method was updated for this study (see Materials and Methods), and we report a large improvement in the size of data set that can be handled. After omitting 2 highly homoplasious sites, we are able to find provably optimal trees for our 127 taxon data set.

Some previous analyses of entire mtDNA data sets have used distance-based analyses, constructing Neighbor-Joining trees (Ingman et al. 2000; Ruiz-Pesini et al. 2004). This is fast computationally, but it is undesirable to use a method that reports a single bifurcating tree (with ties broken arbitrarily) for population data when multifurcations are expected, and as has been noted by Bandelt et al. (1995), homoplasy means that there are typically many equally parsimonious trees. Similar to the median-joining network approach (Bandelt et al. 1995), we want to display

all the equally parsimonious trees in a single graph. Here we accomplish this by summarizing the trees in a consensus network (Holland and Moulton 2003; Holland, Delsuc, et al. 2005) that displays all the edges (splits) in the most parsimonious trees.

Materials and Methods

DNA Amplification and Sequencing

The 19 mitochondrial genomes sequenced in this study are from Taiwan ($n = 2$), Near Oceania ($n = 4$), and Remote Oceania ($n = 13$). A sample from a New Zealander of European maternal descent was also sequenced as a methodological control. The locations and sources of the samples are summarized in table 1. An initial long polymerase chain reaction (PCR) was used to amplify the entire mtDNA in 2 overlapping fragments using the Expand Long Template System (Roche Applied Science, Mannheim, Germany) following the manufacturer's instructions; using primer sets 1F,11R and 11F,1R described by Rieder et al. (1998). Twelve ~2,000-bp internal fragments were subsequently amplified from the large fragments, again using combinations of primers from Rieder et al. (1998). PCR products were purified either by digestion with Exo/Sap or by filtration with PCR Cleanup Plates (Millipore, Billerica, MA) and sequenced in forward and reverse directions using the complete set of 24 overlapping primers. BigDye Terminator chemistry (version 3.1, Applied Biosystems, Foster City, CA) was used to generate sequencing products, and capillary separation was performed on an ABI3730 Genetic Analyzer (Applied Biosystems) by the Allan Wilson Centre Genome Service (Palmerston North, New Zealand). Electropherograms were edited and sequences assembled using Sequencher (Version 4.2.2, Gene Codes Corporation, Ann Arbor, MI).

Data Sets

The 20 sequences from this study were manually aligned using SE-AL (Rambaut 1996) with sequences from previous studies that included Oceanic and East Asian samples (Ingman et al. 2000; Maca-Meyer et al. 2001; Ingman and Gyllensten 2003; Kong et al. 2003; Tanaka et al. 2004; Friedlaender et al. 2005; Macaulay et al. 2005; Merriwether et al. 2005; Starikovskaya et al. 2005; Thangaraj et al. 2005; Trejaut et al. 2005; Kivisild et al. 2006). A 9-bp deletion of 1 copy of a tandem repeat in an intergenic region at nt8270–nt8294 was further encoded in the data set by adding a transition where it occurred.

From this alignment, a subset of 137 mtDNA sequences was selected for the Pacific phylogenetic reconstruction. This Oceanic data set contained an African L3 sequence, (AF347014; Ingman et al. 2000); all sequences from Taiwan, Island Southeast Asia, Oceania, and Australia currently available on public databases; and 19 of the 20 sequences generated by this study (details given in table S1, Supplementary Material online). The noncoding control region (nt16024–nt576) was excluded from this data set reducing the number of unique haplotypes from 137 to 127.

Separate data sets were constructed for haplogroups N/R/B4a, N/R/B5a, M/M7 with M/M22, M/M28 with M/M27, N/R/P with N/R/21, and M/Q with M/M29, each containing the same African L3 sequence as the Oceanic data set. Whereas the P, Q, and M28 haplogroups are autochthonous to Oceania, M7bc, B4a, and B5a mtDNA haplotypes are found in populations outside of this region. Relevant mainland East Asian and Japanese sequences from the original alignment were included in these subsets. The number of haplotypes in these data sets ranged from 9 for B5a to 46 for B4a and the entire mtDNA sequence was analyzed, including the noncoding control region. As the Kivisild et al. (2006) sequences do not include the control region, these were not included in the parsimony analyses. All data sets are available from http://awcmee.massey.ac.nz/downloads.htm.

Phylogenetic Analysis

The MMS (Holland, Huber, et al. 2005) was used to determine optimality of the set of most parsimonious trees found by heuristic analysis. For each data set, the upper bound on the parsimony score was found by heuristic searches carried out using PAUP* 4.0b10 (Swofford 2003) excluding gapped characters (branch swapping = Tree Bisection-Reconnection, stepwise addition = simple). The number of steps required for parsimony-informative characters was calculated using PAUP* and averaged over the sets of most parsimonious trees. A lower bound on the parsimony score was calculated using the MinMax Squeeze program. Exact search is not feasible for such a large data set, but the most parsimonious trees found by heuristic search can be proved optimal if the upper bound found by the PAUP search meets the lower bound found using the MinMax Squeeze program. There is no guarantee that the set of most parsimonious trees found by heuristic search are the only optimal trees, other trees with equal score may exist. In order to be effective on the large data set analyzed here, the MinMax Squeeze software has been improved in 2 main ways from the implementation described in Holland, Huber, et al. 2005. The program was rewritten in the C++ language for improved speed. It uses a new heuristic search routine for finding partitions that takes account of information in the tree used to generate the upper bound; this works by avoiding changes to those parts of the partition where the parsimony score found on the tree and lower bound already agree. The tree used to guide the heuristic search for a good partition can be binary or multifurcating, labels at internal nodes are also allowed. The updated MinMax Squeeze program is available from http://awcmee.massey.ac.nz/downloads.htm.

Consensus networks (Holland and Moulton 2003; Holland, Delsuc, et al. 2005) of the sets of most parsimonious trees found for each data set were constructed using a Python script (available from B.R.H.) in combination with Spectronet 1.27 (Huber et al. 2002). All splits (edges) in the equally parsimonious trees are shown in the consensus networks. Sequencher was used to generate lists of substitutions for each sequence compared with the rCRS (Andrews et al. 1999) (with historic base numbering retained by maintaining the 3107C insertion). Base-labeled phylogenies of the haplogroup subsets were reconstructed from the consensus networks or trees, weighting coding-region changes over control-region polymorphisms when resolving conflicts.

Changes in protein-coding genes were determined as synonymous or nonsynonymous using the MitoAnalyzer tool (2000) (http://www.cstl.nist.gov/biotech/strbase/mitoanalyzer.html).

Molecular Dating

Estimations of the TMRCA were calculated using 3 different rates from previously published studies. The first is calculated solely from synonymous substitutions in the protein-coding genes (Kivisild et al. 2006), whereas the second is derived using all of the mtDNA sequence excluding the control region (Mishmar et al. 2003). Both of these rates are calibrated by comparison to chimpanzee sequences, estimating the most recent common ancestor (MRCA) of human and chimpanzee mtDNA at 6.5 Myr. The third rate is calculated for transitions over a portion of the control region, from nt16090 to nt16365. It is estimated from Native American Eskimo and Na-Dene sequences and calibrated with a date of expansion related to the end of the Younger Dryas glacial relapse (Forster et al. 1996). For each of the 3 rates, the number of relevant substitutions from the vertex of interest to its descendants was averaged (the rho statistic) and the variance (σ^2) calculated as described in Saillard et al. (2000).

Results

Oceanic Data Set

The initial heuristic search on the Oceanic data set excluding the control region found 582,624 most parsimonious trees and provided an upper bound for the MinMax Squeeze (MMS) of 412 (from 282 parsimony-informative characters). The lower bound reached by MMS was 410. Thus, we cannot exclude the possibility that trees requiring 1 fewer mutation (411) or 2 fewer mutations (410) could be found. The consensus network (fig. 2*a*) shows that most of the differences between the trees found involve the branching order from the M and N/R vertices indicating that the tree is not fully resolved at those points. Averaging the number of steps required for each character over all trees identified 5 characters requiring 5 or more steps: nt709 (6.8), nt1598 (6.2), nt1719 (5), nt10398 (5), and nt15924 (5). In several M lineages (M27, M28a, M29, and M42), there is a recurring transition from G to A in the 12S rRNA gene at nt1598. A similar pattern is found at the N/R vertex, where several lineages descending from the N/R vertex appear to have a back mutation at nt10398 from A to G (a nonsynonymous transition at nt10398 in the ND3 gene is 1 of 5 substitutions that define the N macrohaplogroup).

When the 2 sites nt1598 and nt10398 were excluded from the parsimony analysis, 165 trees were found by heuristic search (parsimony-informative characters = 280, search score = 399), and this upper bound of 399 was met by the MMS program, guaranteeing the parsimony score optimal for this reduced data set. Clearly, the 2 excluded sites caused most of the increase in the number of possible trees (165–582,624). The consensus network of the 165 most parsimonious trees is shown in figure 2*b*. Knowing that the number of mutations on this network is minimal allows more definite statements about the early evolution among some lineages.

The exclusion of the 2 characters results in a clear multifurcation at the M vertex and greatly reduces the branching possibilities at the N/R vertex. The consensus network is largely tree-like, with just 1 area of uncertainty within the M/Q haplogroup and another surrounding the branching from N/R of single individuals representing the R12 and R21 haplogroups, an Australian N/R/P sequence, and the remainder of the N/R/P haplogroup. The area of uncertainty in the Q haplogroup involves transitions at nt15172 and nt9254 in the Q3 subhaplogroup (fig. S5, Supplementary Material online); control-region sequence data for the DQI12898 sequence (Kivisild et al. 2006) may help to resolve this part of the network.

The conflicting branching hypotheses at the N/R vertex involve 2 shared coding-region transitions between the R21 and P7 sequences at nt12361 and nt15613, a single shared substitution at nt11404 between the R12 and R21 sequences, and the P-defining substitution at nt15607 that is not found in the R12 and R21 sequences (fig. S4, Supplementary Material online). The link between the Malaysian R21 sequence and the Australian P7 sequence is intriguing and for explanation requires 1) parallel substitutions at 2 sites (nt12361G and nt15613) with a MRCA at the N/R vertex or 2) if it reflects shared ancestry, the P-defining nt15607G transition to have arisen independently in the P7 sequence or reverted to nt15607A in the R21 sequence. Additional sequences from these haplogroups are required to clarify this interesting association.

Haplogroup Subsets

All 6 haplogroup subsets—N/R/B4a (fig. 3), N/R/B5a, M/M7 with M/M22, M/M28 with M/M27, N/R/P with N/R/21, and M/Q with M/M29 (figs. S1–S5, Supplementary Material online)—were proved minimal using the MMS approach. These data sets ranged in size from 9 (N/R/B5a) to 46 (N/R/B4) haplotypes. With these relatively small numbers of sequences, it was possible to include the control region of the mtDNA in the MMS analyses. Three of the analyses (N/R/P with N/R/21, M/M28 with M/M27, and N/R/B5a) resulted in a single most parsimonious tree, and the M/Q with M/M29 analysis found just 2 most parsimonious trees. The N/R/B4 and M/M7 with M/M22 phylogenies were considerably more complex: 1,274 and 74,390 most parsimonious trees were found, respectively. A single tree with branch-labeled nucleotide changes was reconstructed from the minimal tree or consensus network for each of the data sets.

Molecular Dating Estimates

The estimated ages of selected vertices in the haplogroup subset–labeled trees were calculated according to 3 previously described rates based on synonymous changes only (Kivisild et al. 2006), all coding-region changes (Mishmar et al. 2003), and HVR-I changes (Forster et al. 1996) (table 2). The dates estimated from the synonymous changes only are consistently lower than those calculated from all changes in the coding region, whereas the HVR-I estimates are, in several instances, much greater than the coding-region dates. Unless indicated otherwise,

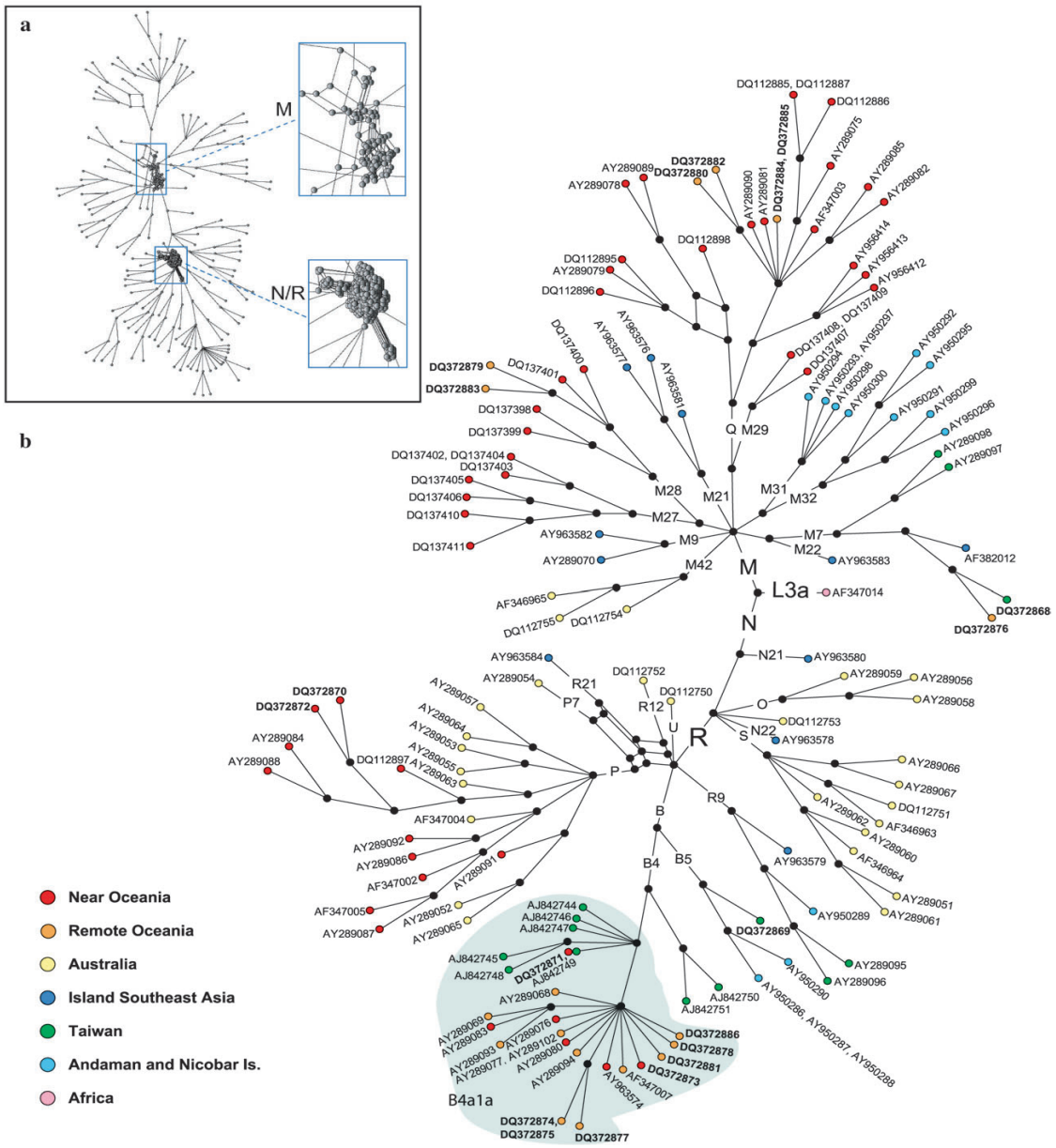


FIG. 2.—Oceania data set consensus networks. (*a*) Consensus of 582,624 most parsimonious trees found by heuristic search; upper bound 412, lower bound 410. The entire coding region (282 parsimony-informative characters) of the mtDNA sequence of 127 haplotypes was analyzed. The 2 major areas of conflict among the trees at the M and N/R vertices are enlarged. (*b*) Consensus network of 165 provably optimal trees found by heuristic search when 2 characters—nt1598 and nt10398—were excluded from the analysis. The MMS guarantees the heuristic search score of 399 to be minimal. Sequences reported here are shown in bold; haplogroups are labeled according to existing nomenclature. The N/R/B4a1a haplogroup is highlighted in green.

we refer to the synonymous substitution rate estimates in the following discussion as these changes are more likely to be selectively neutral than others in the coding region (Penny 2005; Kivisild et al. 2006).

Discussion

MMS Approach to Population Analyses

This analysis demonstrates the effectiveness of the MMS for analyzing population data sets. In most cases,

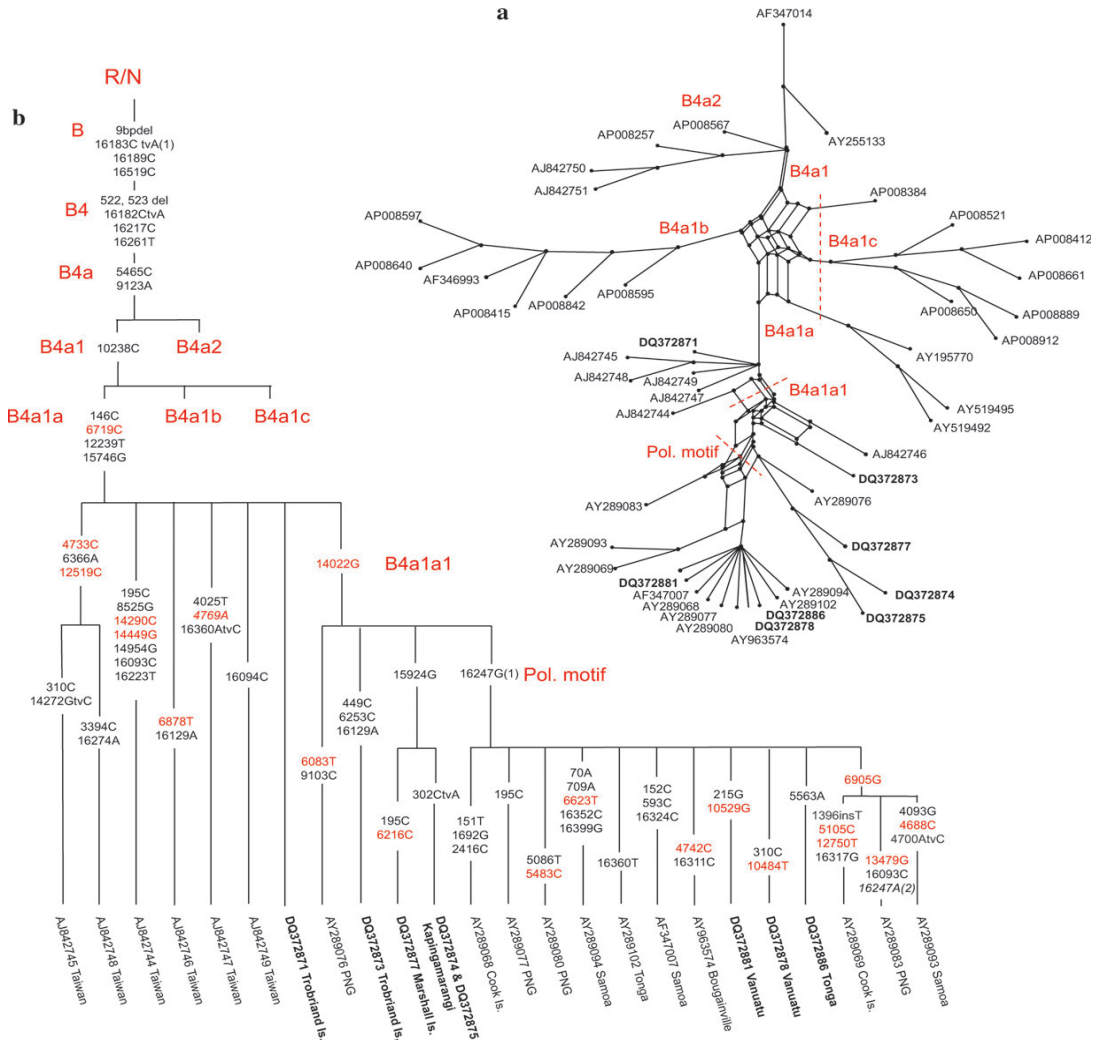


FIG. 3.—B4a consensus network and B4a1a base-labeled phylogeny. (*a*) The consensus network of 1,274 provably optimal trees (57 parsimony-informative characters over complete mtDNA sequence, parsimony score 87) found by heuristic search on the N/R/B4a haplogroup data set of 47 sequences including L3 outgroup (AF347014). Sequences in B4a2 are from Taiwan and Japan, B4a1b sequences are from Japan and Korea, and B4a1c sequences are from Japan and Siberia. All B4a1a sequences are from Taiwan and Oceania. Sequences from this study are shown in bold type. (*b*) A base-labeled phylogeny reconstructed from the consensus network for B4a1a sequences. Length variations in the poly-C region from nt303 to nt315 are not shown. Substitutions are transitions to the base shown, unless marked "tvN," where N is the base in the rCRS. Synonymous substitutions are shown in red type, and sites that change more than once within a lineage in the labeled phylogeny are followed by the number of the change in brackets. When a substitution results in the same nucleotide as in the rCRS, it is shown in italics. The polymorphisms relative to the rCRS at the N/R vertex are: 73G, 263G, 750G, 1438G, 2706G, 3106del, 4769G, 7028T, 8860G, 11719A, 14766T, and 15326G. The conflicts between the trees seen in the consensus network have been resolved here by invoking a reversion at nt16247 from G to A in sequence AY289093. Other positions causing conflict in the phylogeny shown are nt16129 (AJ842746 and DQ372873) and nt16093 (AJ842744 and AY289083).

it was able to prove that a set of equally parsimonious trees was optimal under the parsimony criterion. Holland, Huber, et al. (2005) found that the performance of the MMS at finding the minimum bound was affected by the degree of homoplasy in the data set, and our results support this. Although the entire coding-region data set was not guaranteed optimal, exclusion of 2 of the 5 characters that required a large number of steps in the trees (nt1598 and nt10398) resulted in optimality being provable. For "tip-labeled binary trees," there are $(2n - 5)!!$ trees (Semple and Steel 2003), where n is the number of taxa and the double factorial notation (!!) is multiplying by every second number $(1\times3\times5\cdots2n-5)$. For 127 taxa (the number of unique sequences in the Oceanic data set), there are $\approx 4 \times 10^{245}$ trees. This is far more than that for 53 taxa, which until now was the largest data set for which a tree had been

Table 2
TMRCA Estimates

		Synonymous Transitions[a]			Coding Region nt577–nt16022[b]			HVR-I Transitions nt16090–16365[c]			
Vertex	*n*	ρ	σ	ρ ± σ (years)	ρ	σ	ρ ± σ (years)	*n*	ρ	σ	ρ ± σ (years)
N/R/B4											
B4a1a	25	1.60	0.26	10,822 ± 1,759	2.40	0.31	12,331 ± 1,593	25	1.04	0.20	20,987 ± 4,036
B4a1a1	18	0.78	0.21	5,276 ± 1,429	1.50	0.30	7,707 ± 1,541	18	1.17	0.25	23,611 ± 5,045
'Pol. Motif'	13	0.92	0.27	6,223 ± 1,826	1.54	0.34	7,913 ± 1,747	13	0.54	0.20	10,897 ± 4,036
M/Q											
Q	22	4.95	0.56	33,482 ± 3,788	8.59	0.72	44,135 ± 3,699	16	3.38	0.46	68,208 ± 12,108
Q1 and Q2	16	5.13	0.70	34,699 ± 4,735	7.63	0.83	39,203 ± 4,265	13	3.77	0.54	76,078 ± 10,897
Q1	13	3.15	0.65	21,307 ± 4,387	5.15	0.80	26,460 ± 4,110	10	1.70	0.41	34,306 ± 8,274
Q2	3	3.00	1.00	20,292 ± 6,764	4.67	1.25	23,994 ± 6,423	3	0.33	0.33	6, 659 ± 6,659
N/R/P											
P2	5	1.60	0.57	10,822 ± 3,855	3.60	0.85	18,497 ± 4,367	4	1.50	0.61	30,270 ± 12,310
M/M28											
M28	6	3.00	0.71	20,292 ± 4,802	6.33	1.03	32,524 ± 5,292	6	1.50	0.50	30,270 ± 10, 090
M28a	4	1.50	0.61	10,146 ± 4,126	3.00	0.87	15,414 ± 4,470	4	0.50	0.35	10,090 ± 7,063
N/R/B5											
B5a	10	4.10	0.74	27,732 ± 5,005	5.70	0.88	29,287 ± 4,521	10	1.00	0.37	20,180 ± 7,467

[a] One synonymous transition per 6,764 years (Kivisild et al. 2006).
[b] One substitution per 5,138 years (Mishmar et al. 2003).
[c] One transition per 20,180 years (Forster et al. 1996).

proven minimal (Holland, Huber, et al. 2005); here there were $\approx 3 \times 10^{80}$ tip-labeled binary trees. Thus, we have now shown a major increase in the power of the MinMax Squeeze program.

The use of consensus networks allows for a concise summary of all of the edges contained within the equally parsimonious trees—in particular, highlighting those areas where the trees disagree. We find that combining the MMS and consensus network approaches provides a useful alternative to direct construction of the median network that is practical for large data sets. One advantage of this approach over median networks is that it is not necessary to preprocess the alignment by recoding sites with more than 2 states nor is it necessary to remove sites with ambiguous characters. An improvement to the consensus network method that would make the graphs more representative of population data would be to allow input trees with labeled internal nodes and, hence, to produce networks with internal labels.

Monophyly of Previously Described M Haplogroups

Four of the 7 branches from the M vertex in the Oceanic consensus network combine haplogroups that have been previously reported as direct descendants of M. In 3 of these instances, a single shared character accounts for the link between the named haplogroups. M31 and M32 are found in samples from the Andaman Islands, and the basal link here is a transition from A to G at base nt1524 in the 12S rRNA gene. All of the individuals in M31 and 1 subgroup of 2 of the 5 M32 individuals have this transition. As this variant is found only in the 7 Andaman samples (from 2,423 global sequences in mtDB—Human Mitochondrial Genome Database [Ingman and Gyllensten 2006], 1/02/06), it seems probable that it is a basal polymorphism for a single M haplogroup encompassing all Andaman sequences.

Haplogroups Q and M29 share a synonymous transition in the ND5 gene at nt13500. Merriwether et al. (2005) have suggested that more sequences, particularly from M29, are required to assess whether this is in fact a basal change or independently acquired in each haplogroup as the transition at nt13500 occurs in parallel in several other global lineages. Similarly, as the transition linking M27 and M28 (at nt1719 in the 16S rRNA gene) has arisen several times independently (e.g., in L3e, M/M8/C, M/D, N/R/B4b, and N/R/P), it does not provide strong support for an ancestral link between M27 and M28.

The fourth grouping of 2 M haplogroups previously described independently is between the M7 sequences and M22, represented by a single sequence from Malaysia. These are grouped by 2 shared polymorphisms: a synonymous transition from A to G at nt5351 in the ND2 gene and a nonsynonymous transition from A to G at nt15236 in the cytochrome B gene. Both substitutions are present in the M22 sequence, whereas the nt5231 transition is found in the M7b Taiwanese sequences and the nt15236 polymorphism in the Micronesian and Taiwanese M7c sequences. When other M7 sequences from outside of the Pacific are examined, and the control region included in the analysis (fig. S2, Supplementary Material online), it seems likely that these shared polymorphisms arose independently in M22 and M7.

Autochthonous Haplotypes

There is marked geographic clustering within the Oceanic consensus network (fig. 2*b*). The N/R/P haplogroup contains haplotypes from Australia and Near Oceania, haplogroups M/M42, N/O, and N/S are currently found only in Australia, and M/Q, M/M29, M/M27, and M/M28 are present only in Oceania. Haplogroups M/M31 and M/M32 are found in samples from the Andaman Islands and M/M22, M/M21, N/N21, N/N22, and N/R/R21 only in Malaysian

Aboriginal populations. By contrast, N/R/B4, N/R/B5, N/R/R9, N/R/U, M/M7, and M/M9 haplotypes are found in populations outside of the area covered in this analysis, and their presence in Oceanic populations may represent later migrations into areas initially settled by the ancestors of the present-day carriers of the autochthonous haplotypes.

Eight of the sequences reported here belong to the autochthonous Oceanic haplogroups M/Q, M/M28, and N/R/P. Four M/Q haplotypes are from Polynesia and Vanuatu and are the first reported from Remote Oceania. They form 2 geographic subgroups within the Q1 subclade (fig. S5, Supplementary Material online). The Vanuatu and Polynesian Q1 sequences appear to have diverged well before the settlement of Remote Oceania as they are not closely related to each other or to the other Q1 sequences available at present from Papua New Guinea and Bougainville; the age estimate of the Q1 vertex is 22,862 ± 4,464 years. The Polynesian sequences contain an HVR-I signature, which has been reported from throughout Polynesia; tracking a likely geographic source in Near Oceania of this Q1 variant will require more sampling from the region.

The 2 M/M28 sequences from Remote Oceania (DQ372879 and DQ372883) provide a closer link to known lineages in Near Oceania (fig. S3, Supplementary Material online); the TMRCA of these Vanuatu sequences and 2 from New Britain is 10,146 ± 4,126 years. Two sequences from the Trobriand Islands in Near Oceania (DQ372870 and DQ372872) fall within the N/R/P/P2 subhaplogroup (fig. S4, Supplementary Material online). The P sequences in general are very diverse in HVR-I haplotypes: there is only a single coding-region substitution defining the haplogroup and consequently each of the 7 subhaplogroups has distinct HVR-I sequences. It is interesting to note that the 2 sequences from the Trobriand Islands have an identical HVR-I haplotype to the N/R/HV/H reference sequence (Andrews et al. 1999); if typed to haplogroup solely by HVR-I sequencing, these sequences could be misinterpreted as European-derived haplotypes.

Later mtDNA Arrivals in Oceania

The 2 sequences from Taiwan described here (DQ372868 and DQ372869) belong to haplogroups M/M7c and N/R/B5a, both of which have been reported from HVR-I data to be present in Oceania and Island Southeast Asia. A Micronesian sample from the Marshall Islands (DQ372876) is closely related to the Taiwanese M7c sequence, and these together with a sequence from the Philippines and 1 from Mongolia form a subclade of M7c separate from other sequences from China and Japan (fig. S2, Supplementary Material online). The Taiwanese B5a sequence is notable for its distance from the B5a sequences from the Austro-Asiatic language–speaking Nicobarese (TMRCA estimate 27,732 ± 5,005 years); however, HVR-I sequences from aboriginal Taiwanese suggest that B5a lineages more closely related to the Nicobarese haplotypes may also exist in Taiwan (Trejaut et al. 2005).

The remaining 8 new sequences from Oceania belong to the B4a1a subgroup of haplogroup N/R/B4, bringing the total number of haplotypes in this group to 24. B4a1a HVR-I sequences are the most common type found in Polynesia, and an out-of-Taiwan model predicts the observed pattern of shared ancestry between Oceanic and Taiwanese B4a sequences. The complete mtDNA sequences reveal the phylogeny to be considerably more complex than indicated from the HVR-I sequences: several coding-region substitutions occur between the vertices defining the pre-Polynesian motif (16217C and 16261T) and the full Polynesian motif (16217C, 16247G, and 16261T) (fig. 3).

All of the sequences in the B4a1a subclade to date are from Oceania or Taiwan. The expansion of haplotypes from the B4a1a vertex has occurred recently; a sample from the Trobriand Islands (DQ372871) retains the ancestral sequence, whereas another from Taiwan (AJ842749) differs only by a single control-region transition. The TMRCA estimates for the B4a1a1 and Polynesian motif vertices (table 2) reflect the recent divergence of these sequences and highlight the limitations of molecular dating at the tips of human mtDNA phylogenies and, in particular, estimations based on the noncoding HVR-I. The dates for the B4a1a haplogroup from the entire mtDNA sequences certainly do not exclude the possibility of ancestry of this subhaplogroup in Taiwan, consistent with the out-of-Taiwan model. However, the current lack of entire mtDNA B4a1a sequences from mainland and Island Southeast Asia leaves an important gap in our understanding of the prehistory of this haplogroup.

The haplotypes present in Oceanic populations reinforce the orthodox model of recent settlement of Remote Oceania from Near Oceania, with the maternal lineages in Remote Oceania descending from both the Pleistocene and probable Holocene-era settlers of Near Oceania. The MMS approach is an effective means of analyzing large intraspecific data sets such as this, making it possible to prove that trees found by heuristic search are minimal, and the use of the consensus network method to display all minimal trees found in one graph provides a clear overview of the results, which is easy to interpret. The mitochondrial genome sequences refine existing phylogenies inferred from control-region haplotypes and provide a framework for future investigations: coding-region polymorphisms can be targeted in future studies aimed at resolving the histories of each haplogroup without the need for entire mtDNA sequencing.

Supplementary Materials

Details of the Oceanic data set and results of the haplogroup analyses provided as supplementary table S1 and supplementary figures S1–S5 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Note Added in Proof

Several of the sequences from Andaman Islanders (Thangaraj et al. 2005) have been recently updated (AY950291.2–AY9650300.2, revised 1/06/06). Among the revisions are changes in the sequences at nt1594, which result in the M31 and M32 haplogroups branching independently from the M vertex in contrast to the shared descent shown in figure 2.

Acknowledgments

We are grateful to P.A. McLenachan, W. Schievenhovel, and J. Clegg for providing samples for this project. Thanks are also due to Glenn Conner for implementing the changes to the MMS program and to Lisa Matisoo-Smith for valuable discussions and guidance on issues in Pacific prehistory and scholarship support to M.J.P. This work was supported by a grant from the Marsden Fund.

Literature Cited

Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23:147.

Bandelt HJ, Forster P, Sykes BC, Richards MB. 1995. Mitochondrial portraits of human populations using median networks. Genetics 141:743–53.

Bellwood P. 1991. The Austronesian dispersal and the origin of languages. Sci Am 265:70–5.

Bellwood P. 2001. Early agriculturalist population diasporas? Farming, languages, and genes. Annu Rev Anthropol 30: 181–207.

Cox MP. 2005. Indonesian mitochondrial DNA and its opposition to a Pleistocene era origin of proto-Polynesians in island southeast Asia. Hum Biol 77:179–88.

Diamond J. 1988. Express train to Polynesia. Nature 336:307–8.

Forster P, Harding R, Torroni A, Bandelt HJ. 1996. Origin and evolution of Native American mtDNA variation: a reappraisal. Am J Hum Genet 59:935–45.

Friedlaender J, Schurr T, Gentz F, et al. (13 co-authors). 2005. Expanding southwest pacific mitochondrial haplogroups P and Q. Mol Biol Evol 22:1506–17.

Friedlaender JS, Gentz F, Green K, Merriwether DA. 2002. A cautionary tale on ancient migration detection: mitochondrial DNA variation in Santa Cruz Islands, Solomon Islands. Hum Biol 74:453–71.

Gray RD, Jordan FM. 2000. Language trees support the express-train sequence of Austronesian expansion. Nature 405:1052–5.

Green RC. 2003. The Lapita horizon and traditions—signature for one set of Oceanic migrations. In: Sand C, editor. Pacific archaeology: assessments and anniversary of the first Lapita excavation (July 1952). Volume 15. New Caledonia, France: Le Cahiers de l'Archeologie en Nouvelle-Caledonie, p 95–120.

Hagelberg E, Goldman N, Lio P, Whelan S, Schiefenhovel W, Clegg JB, Bowden DK. 1999. Evidence for mitochondrial DNA recombination in a human population of island Melanesia. Proc R Soc Lond B Biol Sci 266:485–92.

Holland BR, Delsuc F, Moulton V. 2005. Visualizing conflicting evolutionary hypotheses in large collections of trees: using consensus networks to study the origins of placentals and hexapods. Syst Biol 54:66–76.

Holland BR, Huber KT, Penny D, Moulton V. 2005. The MinMax squeeze: guaranteeing a minimal tree for population data. Mol Biol Evol 22:235–42.

Holland BR, Moulton V. 2003. Consensus networks: a method for visualising incompatibilities in collections of trees. Algorithms Bioinformatics Proc 2812:165–76.

Huber KT, Langton M, Penny D, Moulton V, Hendy M. 2002. Spectronet: a package for computing spectra and median networks. Appl Bioinformatics 1:159–61.

Hurles ME, Matisoo-Smith E, Gray RD, Penny D. 2003. Untangling Oceanic settlement: the edge of the knowable. Trends Ecol Evol 18:531–40.

Ingman M, Gyllensten U. 2003. Mitochondrial genome variation and evolutionary history of Australian and New Guinean aborigines. Genome Res 13:1600–6.

Ingman M, Gyllensten U. 2006. mtDB: human mitochondrial genome database, a resource for population genetics and medical sciences. Nucleic Acids Res 34:D749–51.

Ingman M, Kaessmann H, Paabo S, Gyllensten U. 2000. Mitochondrial genome variation and the origin of modern humans. Nature 408:708–13.

Kirch PV. 2000. On the road of the winds: an archaeological history of the Pacific Islands before European contact. Berkeley, CA: University of California Press.

Kivisild T, Shen P, Wall DP, et al. (17 co-authors). 2006. The role of selection in the evolution of human mitochondrial genomes. Genetics 172:373–87.

Kong QP, Yao YG, Sun C, Bandelt HJ, Zhu CL, Zhang YP. 2003. Phylogeny of East Asian mitochondrial DNA lineages inferred from complete sequences. Am J Hum Genet 73: 671–6.

Lum JK, Cann RL, Martinson JJ, Jorde LB. 1998. Mitochondrial and nuclear genetic relationships among Pacific Island and Asian populations. Am J Hum Genet 63:613–24.

Maca-Meyer N, Gonzalez AM, Larruga JM, Flores C, Cabrera VM. 2001. Major genomic mitochondrial lineages delineate early human expansions. BMC Genet 2:13.

Macaulay V, Hill C, Achilli A, et al. (21 co-authors). 2005. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science 308:1034–6.

Melton T, Peterson R, Redd AJ, Saha N, Sofro AS, Martinson J, Stoneking M. 1995. Polynesian genetic affinities with Southeast Asian populations as identified by mtDNA analysis. Am J Hum Genet 57:403–14.

Merriwether DA, Hodgson JA, Friedlaender FR, Allaby R, Cerchio S, Koki G, Friedlaender JS. 2005. Ancient mitochondrial M haplogroups identified in the Southwest Pacific. Proc Natl Acad Sci USA 102:13034–9.

Mishmar D, Ruiz-Pesini E, Golik P, et al. (13 co-authors). 2003. Natural selection shaped regional mtDNA variation in humans. Proc Natl Acad Sci USA 100:171–6.

MitoAnalyzer. 2000. Gaithersburg, MD: National Institute of Standards and Technology.

Murray-McIntosh RP, Scrimshaw BJ, Hatfield PJ, Penny D. 1998. Testing migration patterns and estimating founding population size in Polynesia by using human mtDNA sequences. Proc Natl Acad Sci USA 95:9047–52.

O'Connell JF, Allen J. 2004. Dating the colonization of Sahul (Pleistocene Australia-New Guinea): a review of recent research. J Archaeol Sci 31:835–53.

Oppenheimer S. 2004. The 'express train from Taiwan to Polynesia': on the congruence of proxy lines of evidence. World Archaeol 36:591–600.

Penny D. 2005. Evolutionary biology: relativity for molecular clocks. Nature 436:183–4.

Rambaut A. 1996. Se-Al: sequence alignment editor. Available at: http://evolve.zoo.ox.ac.uk/. Accessed 2005 Jan 01.

Redd AJ, Takezaki N, Sherry ST, McGarvey ST, Sofro AS, Stoneking M. 1995. Evolutionary history of the COII/tRNALys intergenic 9 base pair deletion in human mitochondrial DNAs from the Pacific. Mol Biol Evol 12: 604–15.

Richards M, Oppenheimer S, Sykes B. 1998. mtDNA suggests Polynesian origins in Eastern Indonesia. Am J Hum Genet 63:1234–6.

Rieder MJ, Taylor SL, Tobe VO, Nickerson DA. 1998. Automating the identification of DNA variations using quality-based fluorescence re-sequencing: analysis of the human mitochondrial genome. Nucleic Acids Res 26:967–73.

Ruiz-Pesini E, Mishmar D, Brandon M, Procaccio V, Wallace DC. 2004. Effects of purifying and adaptive selection on regional variation in human mtDNA. Science 303:223–6.

Saillard J, Forster P, Lynnerup N, Bandelt HJ, Norby S. 2000. mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. Am J Hum Genet 67:718–26.

Semple C, Steel M. 2003. Phylogenetics. New York: Oxford University Press.

Starikovskaya EB, Sukernik RI, Derbeneva OA, et al. (11 co-authors). 2005. Mitochondrial DNA diversity in indigenous populations of the southern extent of Siberia, and the origins of Native American haplogroups. Ann Hum Genet 69: 67–89.

Steel M, Penny D. 2004. Two further links between MP and ML under the Poisson model. Appl Math Lett 17:785–90.

Swofford DL. 2003. PAUP* phylogenetic analysis using parsimony (*and other methods). Sunderland, MA: Sinauer Associates.

Sykes B, Leiboff A, Low-Beer J, Tetzner S, Richards M. 1995. The origins of the Polynesians: an interpretation from mitochondrial lineage analysis. Am J Hum Genet 57:1463–75.

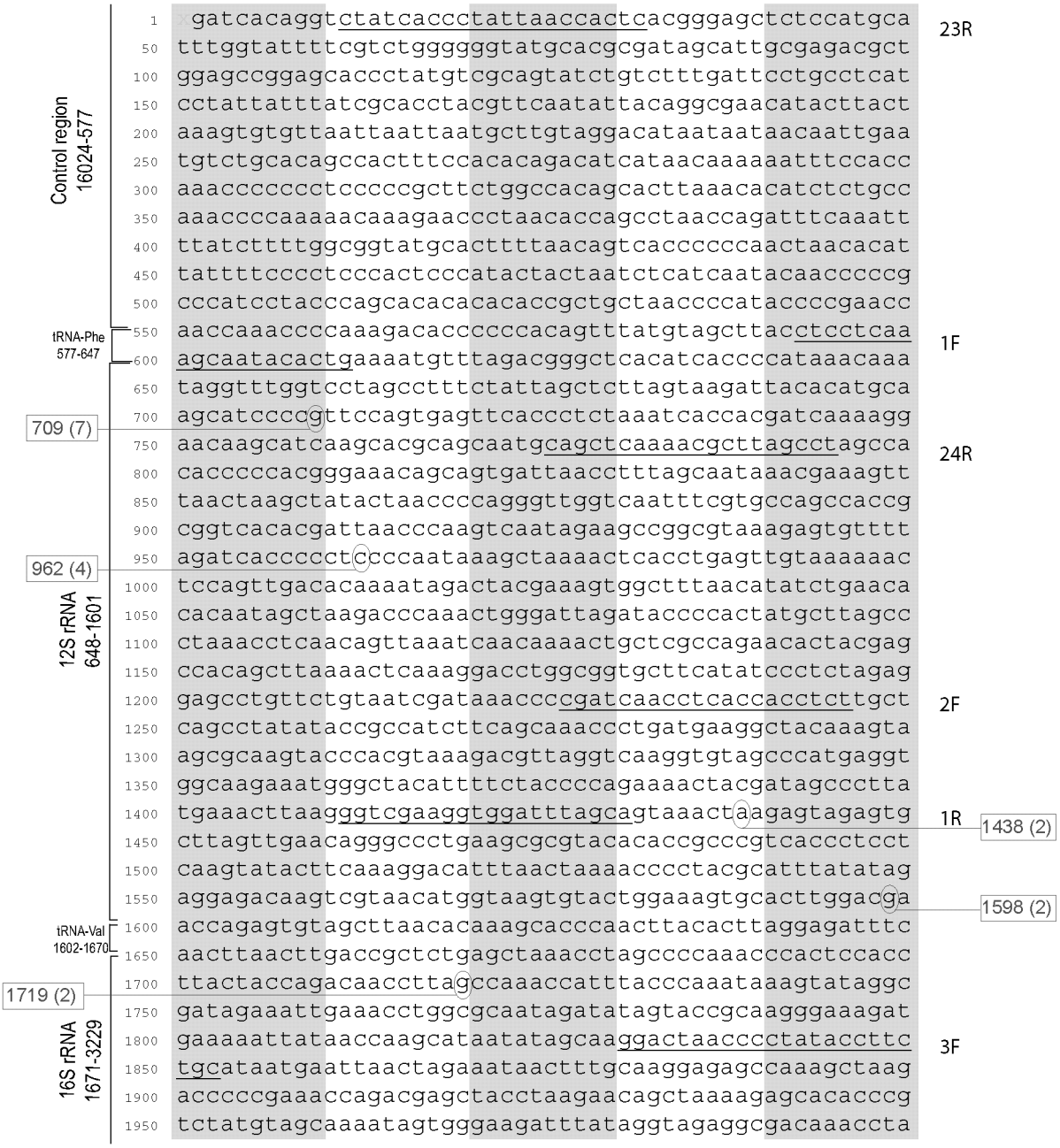
Tanaka M, Cabrera VM, Gonzalez AM, et al. (28 co-authors). 2004. Mitochondrial genome variation in eastern Asia and the peopling of Japan. Genome Res 14:1832–50.

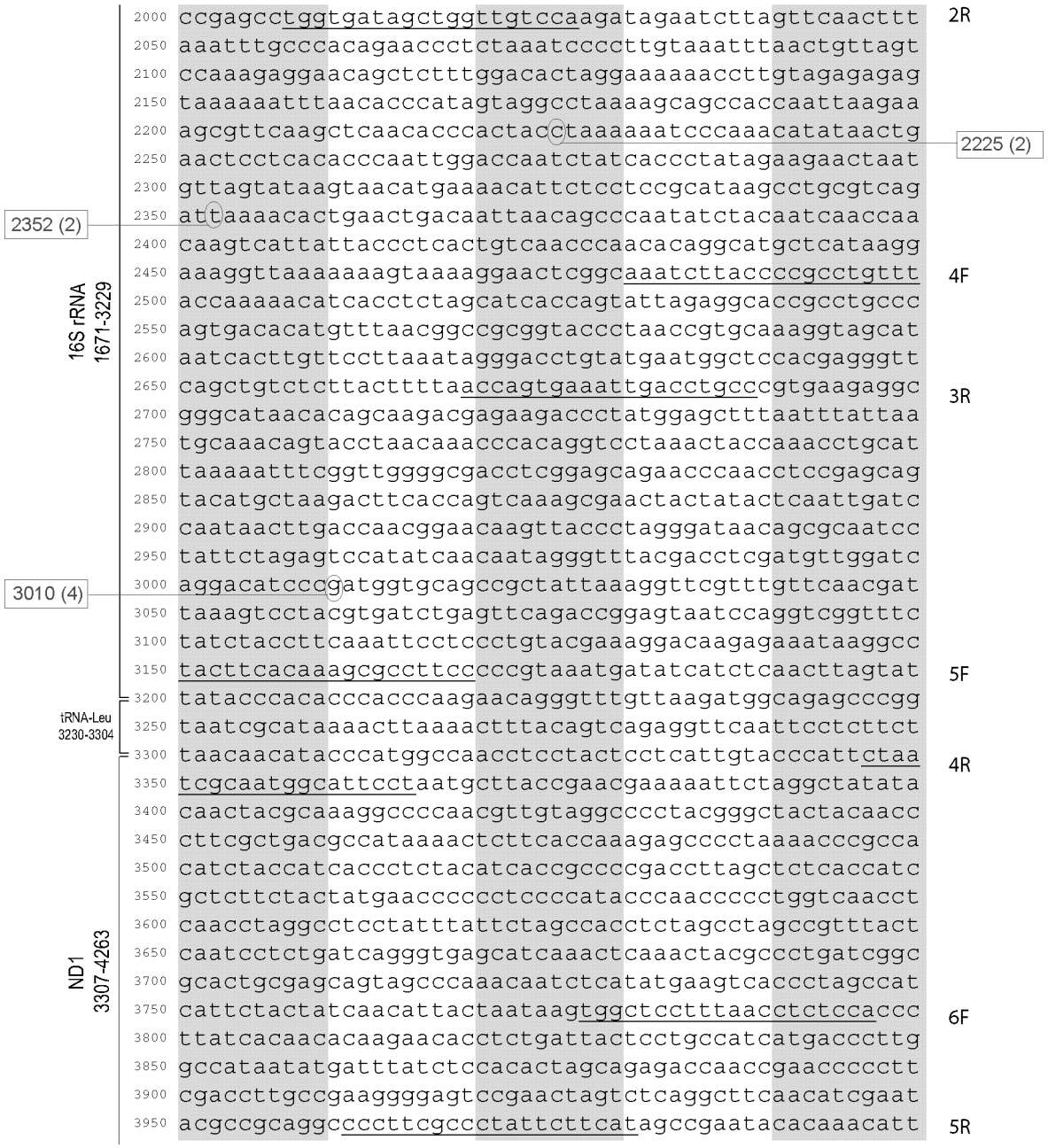
Thangaraj K, Chaubey G, Kivisild T, Reddy AG, Singh VK, Rasalkar AA, Singh L. 2005. Reconstructing the origin of Andaman Islanders. Science 308:996.

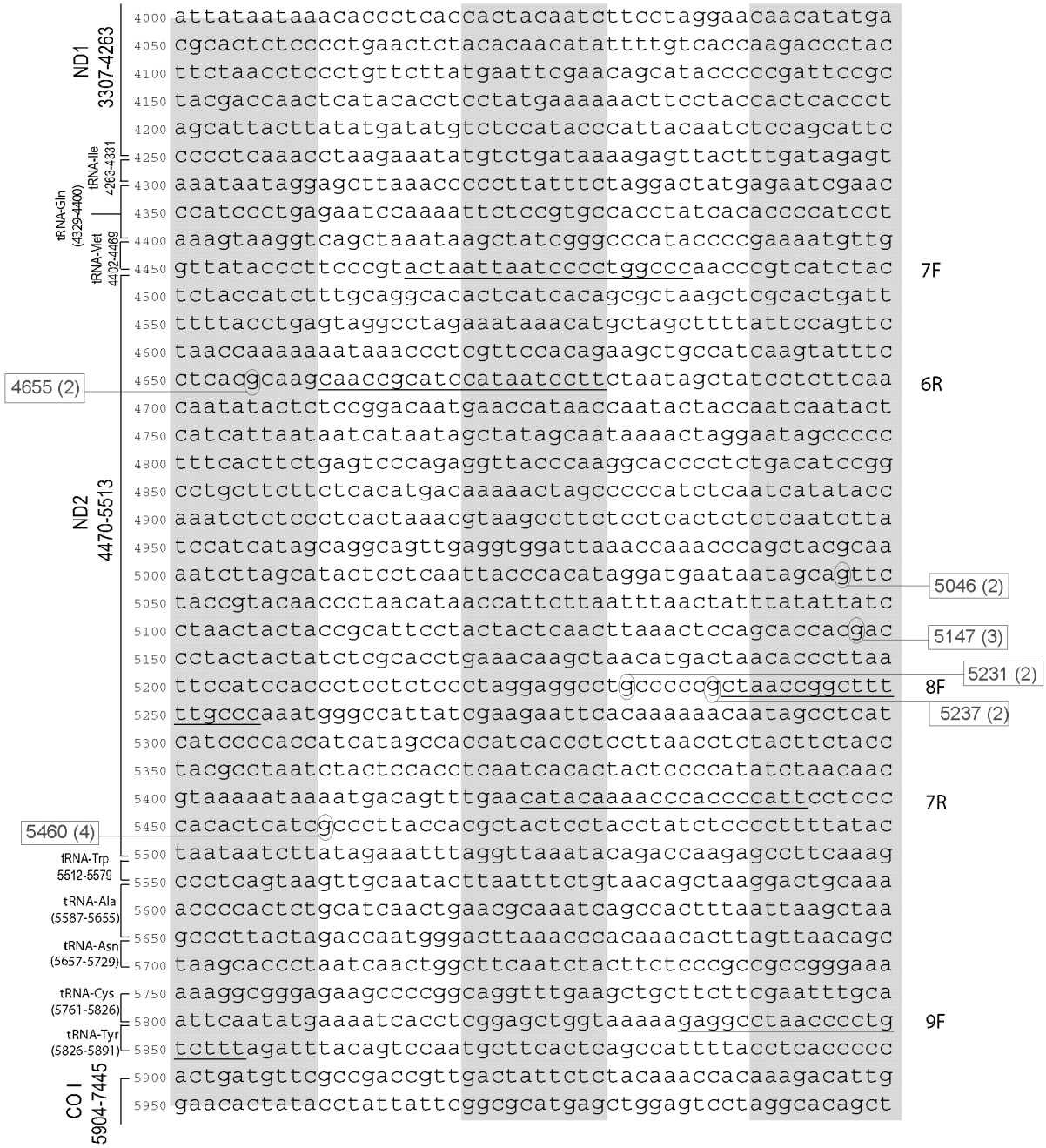
Trejaut JA, Kivisild T, Loo JH, Lee CL, He CL, Hsu C J, Li ZY, Lin M. 2005. Traces of archaic mitochondrial lineages persist in Austronesian-speaking Formosan populations. PLoS Biol 3:e247.

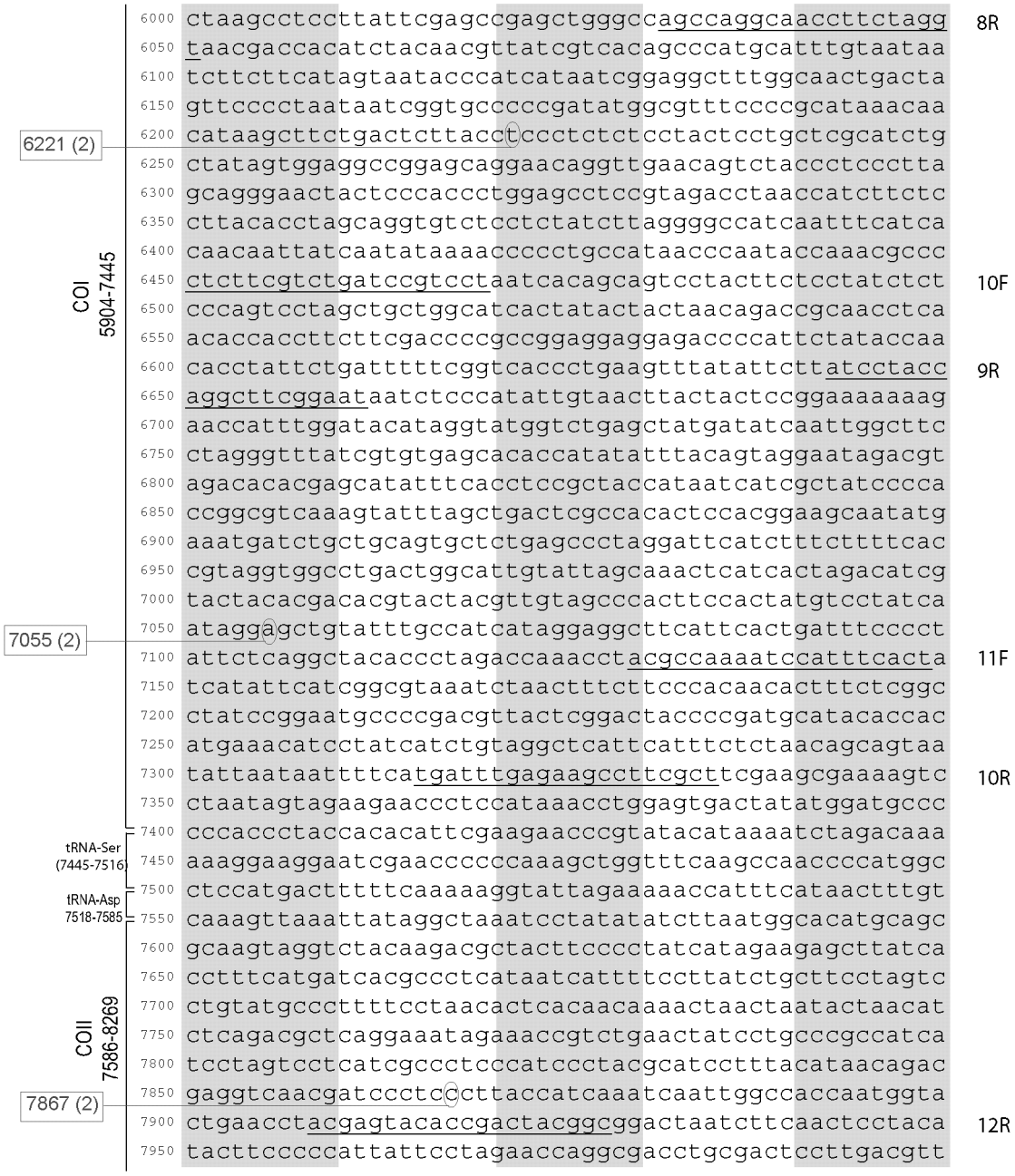
Dan Graur, Associate Editor

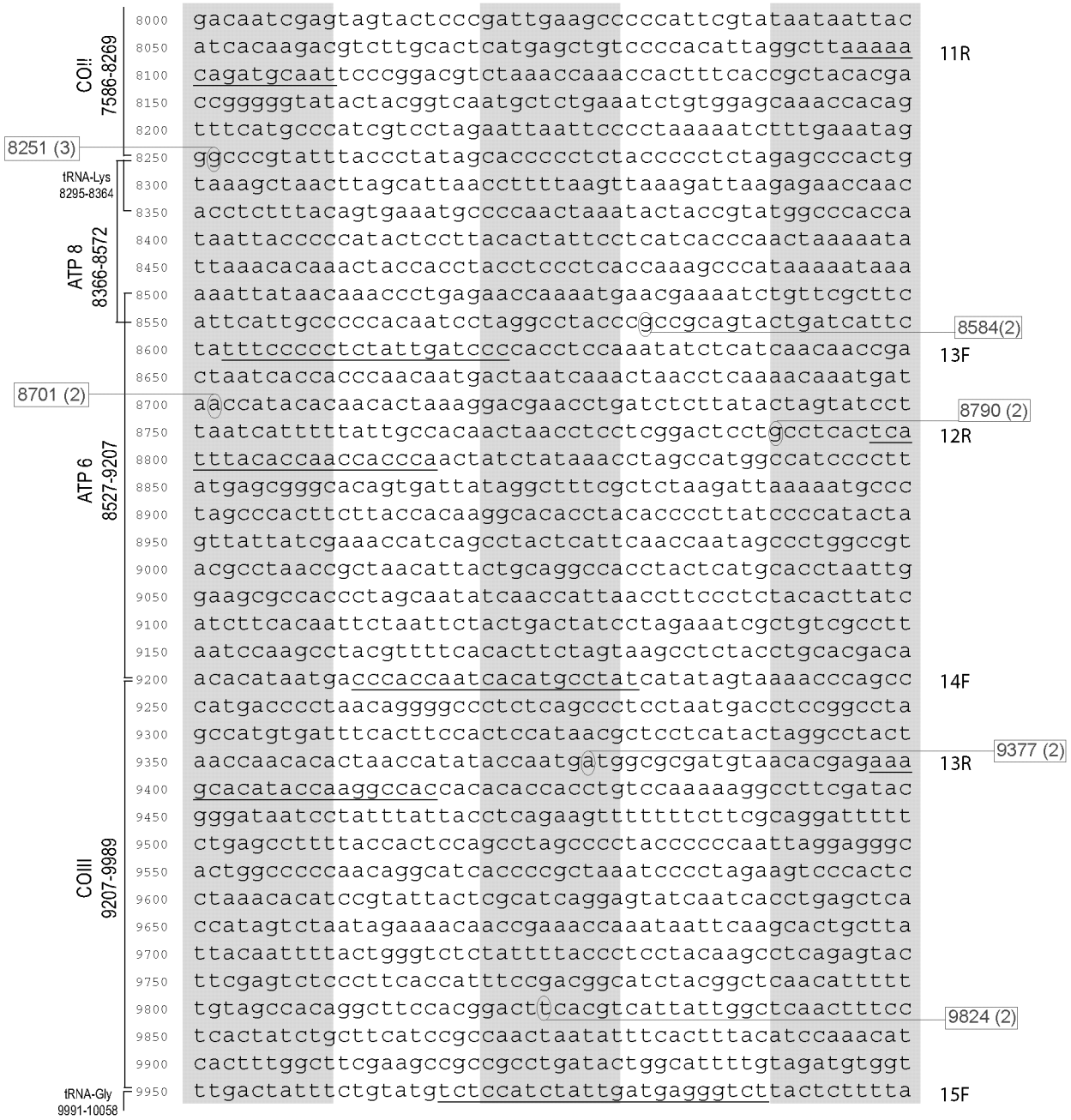
Accepted July 14, 2006

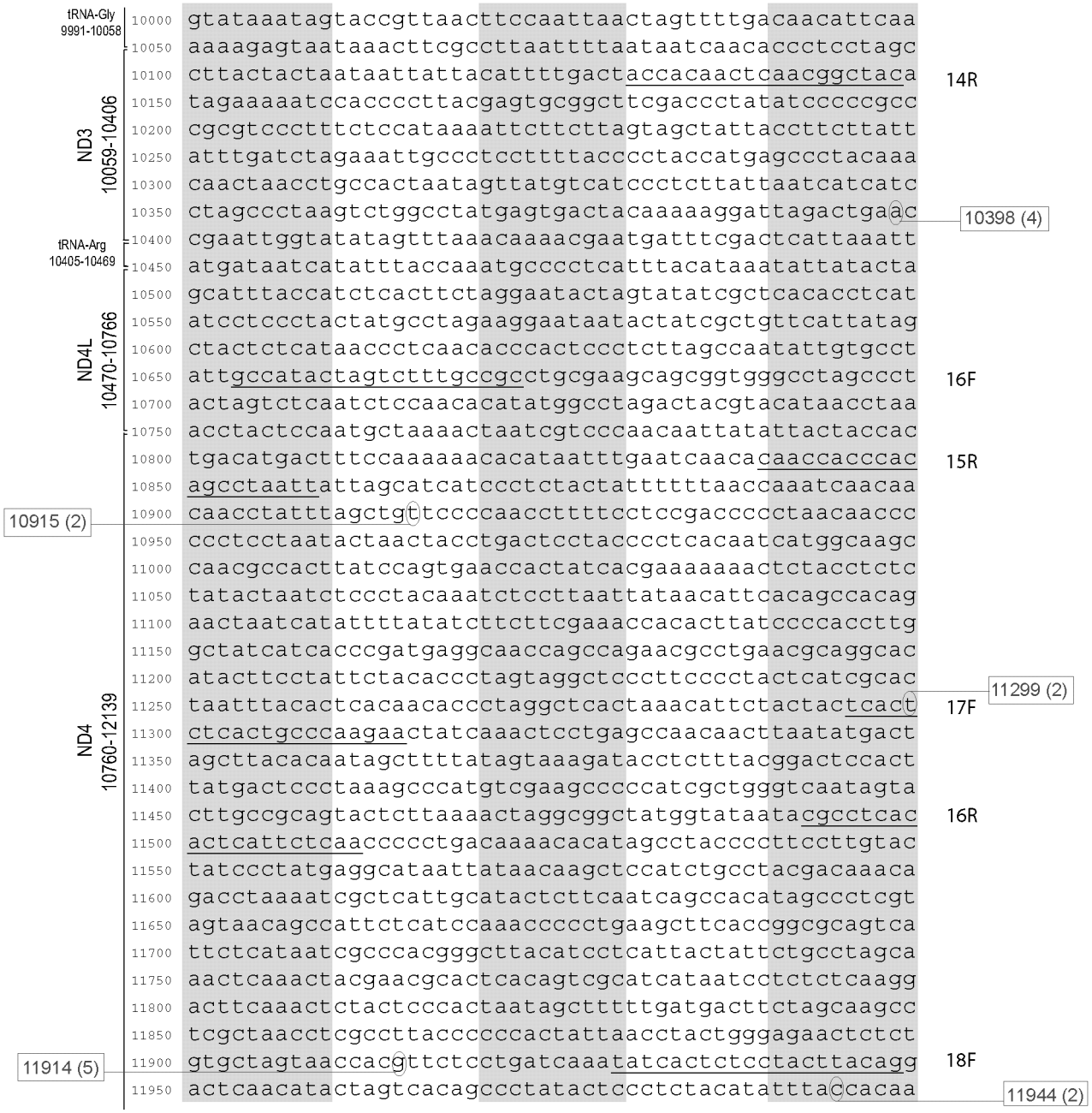

Control region
16024-577
tRNA-Phe
577-647
12S rRNA
648-1601
tRNA-Val
1602-1670
16S rRNA
1671-3229
709 (7)
962 (4)
1719 (2)
1438 (2)
1598 (2)
23R
1F
24R
2F
1R
3F

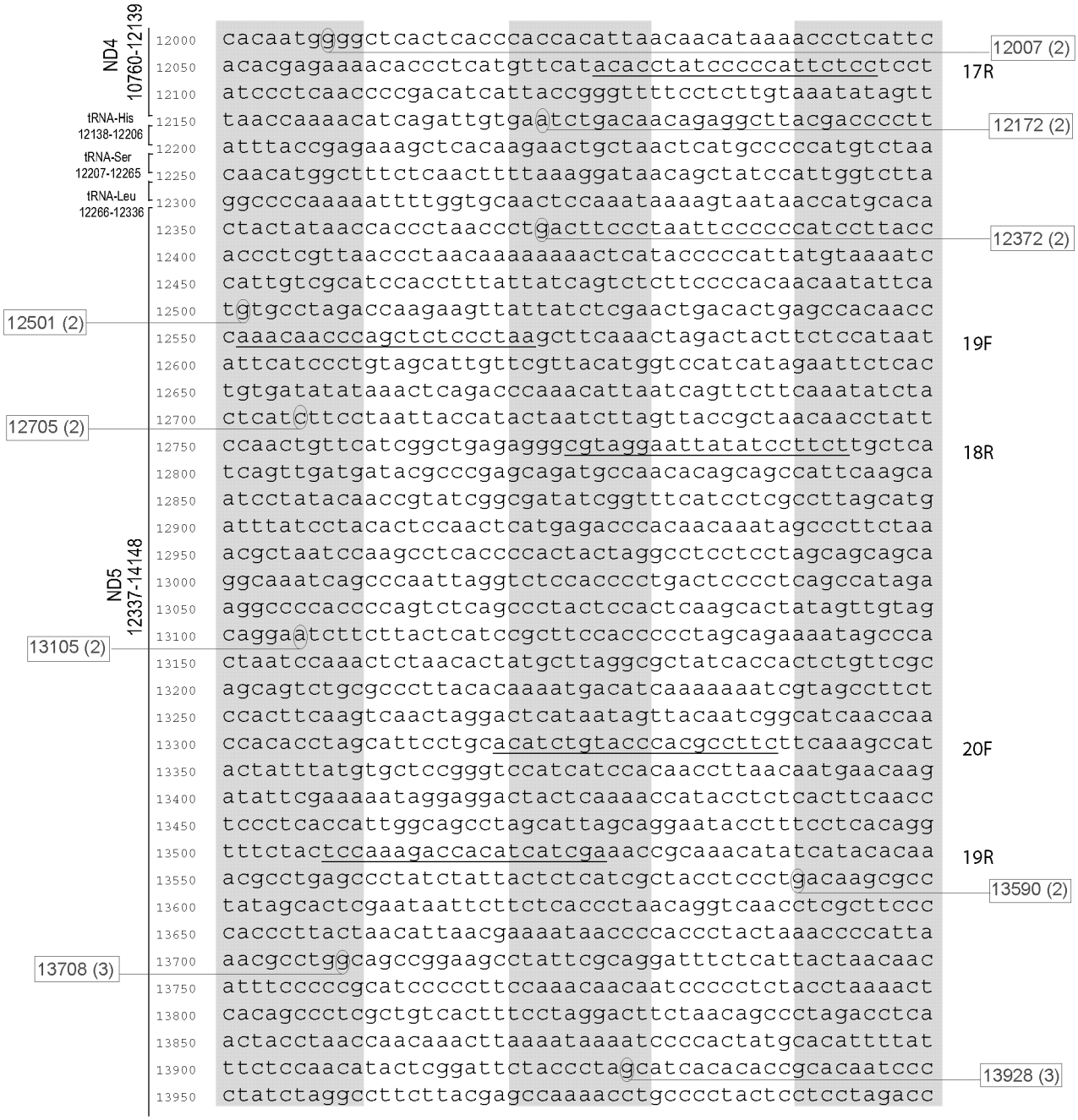

2352 (2)
2225 (2)
3010 (4)
16S rRNA
1671-3229
tRNA-Leu
3230-3304
ND1
3307-4263
2R
4F
3R
5F
4R
6F
5R

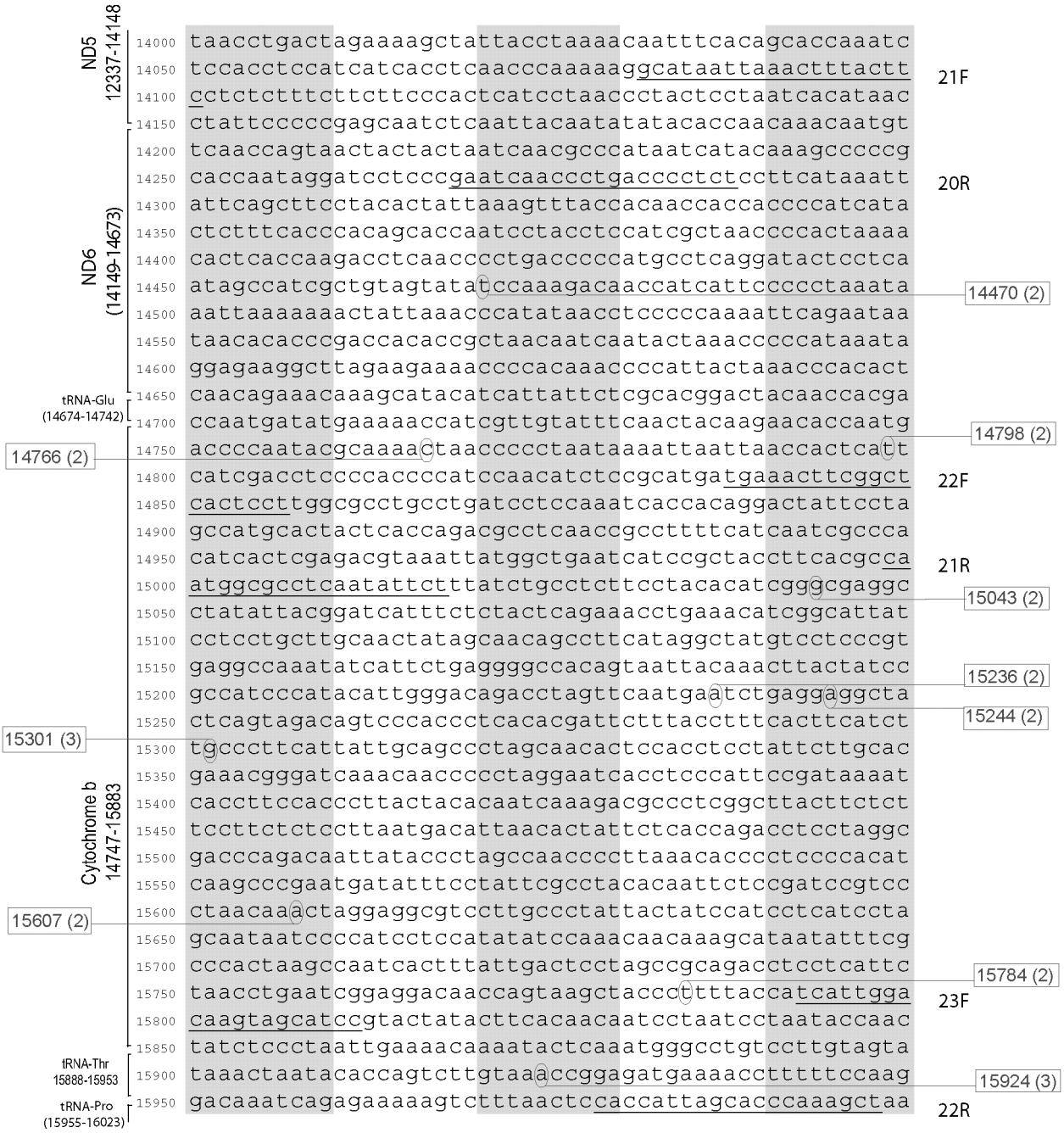

ND1
3307-4263
tRNA-Ile
4263-4331
tRNA-Gln
(4329-4400)
tRNA-Met
4402-4469
ND2
4470-5513
CO I
5904-7445
tRNA-Trp
5512-5579
tRNA-Ala
(5587-5655)
tRNA-Asn
(5657-5729)
tRNA-Cys
(5761-5826)
tRNA-Tyr
(5826-5891)
4655 (2)
5460 (4)
5046 (2)
5147 (3)
5231 (2)
5237 (2)
7F
6R
8F
7R
9F

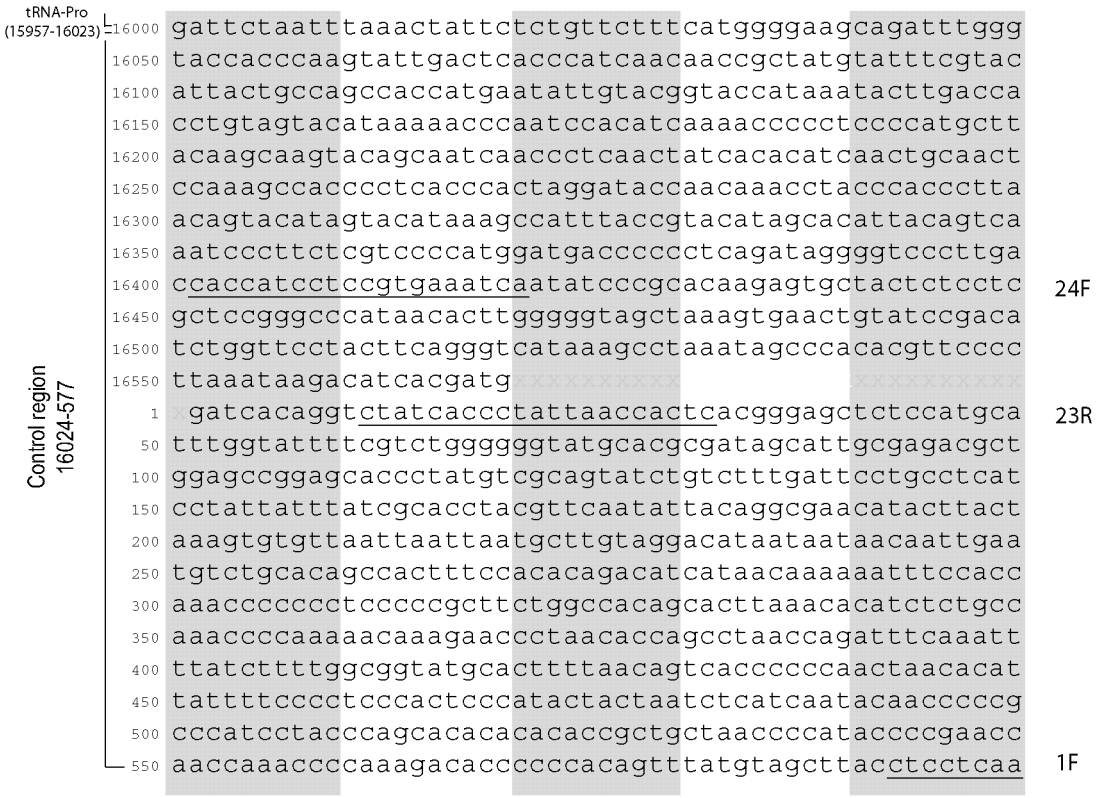

6000 ctaagcctccttattcgagccgagctgggccagccaggcaaccttctagg 8R
6050 taacgaccacatctacaacgttatcgtcacagcccatgcatttgtaataa
6100 tcttcttcatagtaatacccatcataatcggaggctttggcaactgacta
6150 gttcccctaataatcggtgcccccgatatggcgtttccccgcataaacaa
6200 cataagcttctgactcttacctccctctctcctactcctgctcgcatctg
6221 (2)
6250 ctatagtggaggccggagcaggaacaggttgaacagtctaccctccctta
6300 gcagggaactactcccaccctggagcctccgtagacctaaccatcttctc
6350 cttacacctagcaggtgtctcctctatcttaggggccatcaatttcatca
6400 caacaattatcaatataaaaccccctgccataacccaataccaaacgccc
COI 5904-7445
6450 ctcttcgtctgatccgtcctaatcacagcagtcctacttctcctatctct 10F
6500 cccagtcctagctgctggcatcactatactactaacagaccgcaacctca
6550 acaccaccttcttcgaccccgccggaggaggagaccccattctataccaa
6600 cacctattctgatttttcggtcaccctgaagtttatattcttatcctacc 9R
6650 aggcttcggaataatctcccatattgtaacttactactccggaaaaaaag
6700 aaccatttggatacataggtatggtctgagctatgatatcaattggcttc
6750 ctagggtttatcgtgtgagcacaccatatatttacagtaggaatagacgt
6800 agacacacgagcatatttcacctccgctaccataatcatcgctatcccca
6850 ccggcgtcaaagtatttagctgactcgccacactccacggaagcaatatg
6900 aaatgatctgctgcagtgctctgagccctaggattcatctttcttttcac
6950 cgtaggtggcctgactggcattgtattagcaaactcatcactagacatcg
7000 tactacacgacacgtactacgttgtagcccacttccactatgtcctatca
7050 atagggctgtatttgccatcataggaggcttcattcactgatttccccct
7055 (2)
7100 attctcaggctacaccctagaccaaacctacgccaaaatccatttcacta 11F
7150 tcatattcatcggcgtaaatctaactttcttcccacaacactttctcggc
7200 ctatccggaatgccccgacgttactcggactaccccgatgcatacaccac
7250 atgaaacatcctatcatctgtaggctcattcatttctctaacagcagtaa
7300 tattaataatttcatgatttgagaagccttcgcttcgaagcgaaaagtc 10R
7350 ctaatagtagaagaaccctccataaacctggagtgactatatggatgccc
7400 cccaccctaccacacattcgaagaacccgtatacataaaatctagacaaa
tRNA-Ser (7445-7516)
7450 aaaggaaggaatcgaaccccccaaagctggtttcaagccaaccccatggc
7500 ctccatgacttttcaaaaaggtattagaaaaaccatttcataactttgt
tRNA-Asp 7518-7585
7550 caaagttaaattataggctaaatcctatatatcttaatggcacatgcagc
7600 gcaagtaggtctacaagacgctacttcccctatcatagaagagcttatca
7650 cctttcatgatcacgccctcataatcattttccttatctgcttcctagtc
COII 7586-8269
7700 ctgtatgccctttttcctaacactcacaacaaaactaactaatactaacat
7750 ctcagacgctcaggaaatagaaaccgtctgaactatcctgcccgccatca
7800 tcctagtcctcatcgccctcccatccctacgcatcctttacataacagac
7850 gaggtcaacgatccctccccttaccatcaaatcaattggccaccaatggta
7867 (2)
7900 ctgaacctacgagtacaccgactacggcggactaatcttcaactcctaca 12R
7950 tacttcccccattattcctagaaccaggcgacctgcgactccttgacgtt


COII
7586-8269
tRNA-Lys
8295-8364
ATP 8
8366-8572
ATP 6
8527-9207
COIII
9207-9989
tRNA-Gly
9991-10058
8251 (3)
8701 (2)
8584(2)
8790 (2)
9377 (2)
9824 (2)
11R
13F
12R
14F
13R
15F
8000 gacaatcgagtagtactcccgattgaagcccccattcgtataataattac
8050 atcacaagacgtcttgcactcatgagctgtccccacattaggcttaaaaa
8100 cagatgcaattcccggacgtctaaaccaaaccactttcaccgctacacga
8150 ccgggggtatactacggtcaatgctctgaaatctgtggagcaaaccacag
8200 tttcatgcccatcgtcctagaattaattcccctaaaaatctttgaaatag
8250 ggcccgtatttaccctatagcacccccctctaccccctctagagcccactg
8300 taaagctaacttagcattaacctttttaagttaaagattaagagaaccaac
8350 acctctttacagtgaaatgccccaactaaatactaccgtatggcccacca
8400 taattaccccatactccttacactattcctcatcacccaactaaaaata
8450 ttaaacacaaactaccacctacctcccctcaccaaagcccataaaaataaa
8500 aaattataacaaaccctgagaaccaaaatgaacgaaaatctgttcgcttc
8550 attcattgcccccacaatcctaggcctacccgccgcagtactgatcattc
8600 tatttccccctctattgatcccccacctccaaatatctcatcaacaaccga
8650 ctaatcaccacccaacaatgactaatcaaactaacctcaaaacaaatgat
8700 aaccatacacaacactaaaggacgaacctgatctcttatactagtatcct
8750 taatcatttttattgccacaactaacctcctcggactcctgcctcactca
8800 tttacaccaaccacccaactatctataaacctagccatggccatcccctt
8850 atgagcgggcacagtgattataggctttcgctctaagattaaaaatgccc
8900 tagcccacttcttaccacaaggcacacctacacccctttatccccatacta
8950 gttattatcgaaaccatcagcctactcattcaaccaatagccctggccgt
9000 acgcctaaccgctaacattactgcaggccacctactcatgcacctaattg
9050 gaagcgccaccctagcaatatcaaccattaaccttccctctacacttatc
9100 atcttcacaattctaattctactgactatcctagaaatcgctgtcgcctt
9150 aatccaagcctacgttttcacacttctagtaagcctctacctgcacgaca
9200 acacataatgacccaccaatcacatgcctatcatatagtaaaacccagcc
9250 catgacccctaacaggggccctctcagccctcctaatgacctccggccta
9300 gccatgtgatttcacttccactccataacgctcctcatactaggcctact
9350 aaccaacacactaaccatataccaatgatggcgcgatgtaacacgagaaa
9400 gcacataccaaggccaccacacaccacctgtccaaaaaggccttcgatac
9450 gggataatcctatttattacctcagaagtttttttcttcgcaggattttt
9500 ctgagccttttaccactccagcctagcccctaccccccaattaggagggc
9550 actggcccccaacaggcatcaccccgctaaatcccctagaagtcccactc
9600 ctaaacacatccgtattactcgcatcaggagtatcaatcacctgagctca
9650 ccatagtctaatagaaaacaaccgaaaccaaataattcaagcactgctta
9700 ttacaattttactgggtctctatttaccctcctacaagcctcagagtac
9750 ttcgagtctcccttcaccatttccgacggcatctacggctcaacattttt
9800 tgtagccacaggcttccacggactcacgtcattattggctcaactttcc
9850 tcactatctgcttcatccgccaactaatatttcactttacatccaaacat
9900 cactttggcttcgaagccgccgcctgatactggcattttgtagatgtggt
9950 ttgactatttctgtatgtctccatctattgatgagggtcttactctttta


tRNA-Gly
9991-10058
ND3
10059-10406
tRNA-Arg
10405-10469
ND4L
10470-10766
ND4
10760-12139
14R
16F
15R
16R
17F
18F
10398 (4)
10915 (2)
11299 (2)
11914 (5)
11944 (2)


ND4
10760-12139
tRNA-His
12138-12206
tRNA-Ser
12207-12265
tRNA-Leu
12266-12336
ND5
12337-14148
12000 cacaatgggctcactcacccaccacattaacaacataaaaccctcattc
12050 acacgagaaaacaccctcatgttcatacacctatcccccattctcctcct
12100 atccctcaaccccgacatcattaccgggttttcctcttgtaaatatagtt
12150 taaccaaaacatcagattgtgaatctgacaacagaggcttacgacccctt
12200 atttaccgagaaagctcacaagaactgctaactcatgcccccatgtctaa
12250 caacatggctttctcaacttttaaaggataacagctatccattggtctta
12300 ggccccaaaaatttggtgcaactccaaataaaagtaataaccatgcaca
12350 ctactataaccaccctaaccctgacttccctaattccccccatccttacc
12400 accctcgttaaccctaacaaaaaaaactcataccccattatgtaaaatc
12450 cattgtcgcatccacctttattatcagtctcttccccacaacaatattca
12500 tgtgcctagaccaagaagttattatctcgaactgacactgagccacaacc
12550 caaacaacccagctctccctaagcttcaaactagactacttctccataat
12600 attcatccctgtagcattgttcgttacatggtccatcatagaattctcac
12650 tgtgatatataaactcagacccaaacattaatcagttcttcaaatatcta
12700 ctcatcttcctaattaccatactaatcttagttaccgctaacaacctatt
12750 ccaactgttcatcggctgagagggcgtaggaattatatccttcttgctca
12800 tcagttgatgatacgcccgagcagatgccaacacagcagccattcaagca
12850 atcctatacaaccgtatcggcgatatcggtttcatcctcgccttagcatg
12900 atttatcctacactccaactcatgagacccacaacaaatagcccttctaa
12950 acgctaatccaagcctcaccccactactaggcctcctcctagcagcagca
13000 ggcaaatcagcccaattaggtctccacccctgactcccctcagccataga
13050 aggccccaccccagtctcagccctactccactcaagcactatagttgtag
13100 caggaatcttcttactcatccgcttccacccccctagcagaaaatagccca
13150 ctaatccaaactctaacactatgcttaggcgctatcaccactctgttcgc
13200 agcagtctgcgcccttacacaaaatgacatcaaaaaaatcgtagccttct
13250 ccacttcaagtcaactaggactcataatagttacaatcggcatcaaccaa
13300 ccacacctagcattcctgcacatctgtacccacgccttcttcaaagccat
13350 actatttatgtgctccgggtccatcatccacaaccttaacaatgaacaag
13400 atattcgaaaaataggaggactactcaaaaccatacctctcacttcaacc
13450 tccctcaccattggcagcctagcattagcaggaatacctttcctcacagg
13500 tttctactccaaagaccacatcatcgaaaccgcaaacatatcatacacaa
13550 acgcctgagccctatctattactctcatcgctacctccctgacaagcgcc
13600 tatagcactcgaataattcttctcaccctaacaggtcaacctcgcttccc
13650 caccttactaacattaacgaaaataaccccacccctactaaaccccatta
13700 aacgcctggcagccggaagcctattcgcaggatttctcattactaacaac
13750 atttcccccgcatccccccttccaaacaacaatccccctctacctaaaact
13800 cacagccctcgctgtcactttcctaggacttctaacagccctagacctca
13850 actacctaaccaacaaacttaaaataaaatccccactatgcacatttttat
13900 ttctccaacatactcggattctaccctagcatcacacaccgcacaatccc
13950 ctatctaggccttcttacgagccaaaacctgcccctactcctcctagacc
12007 (2)
17R
12172 (2)
12372 (2)
12501 (2)
19F
12705 (2)
18R
13105 (2)
20F
19R
13590 (2)
13708 (3)
13928 (3)


ND5
12337-14148
ND6
(14149-14673)
tRNA-Glu
(14674-14742)
Cytochrome b
14747-15883
tRNA-Thr
15888-15953
tRNA-Pro
(15955-16023)
14000 taacctgactagaaaagctattacctaaaacaatttcacagcaccaaatc
14050 tccacctccatcatcacctcaacccaaaaaggcataattaaactttactt
14100 cctctctttcttcttcccactcatcctaaccctactcctaatcacataac
14150 ctattccccgagcaatctcaattacaatatatacaccaacaaacaatgt
14200 tcaaccagtaactactactaatcaacgcccataatcatacaaagcccccg
14250 caccaataggatcctcccgaatcaaccctgacccctctccttcataaatt
14300 attcagcttcctacactattaaagtttaccacaaccaccaccccatcata
14350 ctctttcacccacagcaccaatcctacctccatcgctaaccccactaaaa
14400 cactcaccaagacctcaaccccctgaccccatgcctcaggatactcctca
14450 atagccatcgctgtagtatatccaaagacaaccatcattccccctaaata
14500 aattaaaaaaactattaaacccatataacctcccccaaaattcagaataa
14550 taacacacccgaccacaccgctaacaatcaatactaaaccccctataaata
14600 ggagaaggcttagaagaaaaccccacaaaccccattactaaacccacact
14650 caacagaaacaaagcatacatcattattctcgcacggactacaaccacga
14700 ccaatgatatgaaaaaccatcgttgtatttcaactacaagaacaccaatg
14750 accccaatacgcaaaactaacccccctaataaaattaattaaccactcatt
14800 catcgacctccccacccccatccaacatctccgcatgatgaaacttcggct
14850 cactccttggcgcctgcctgatcctccaaatcaccacaggactattccta
14900 gccatgcactactcaccagacgcctcaaccgccttttcatcaatcgccca
14950 catcactcgagacgtaaattatggctgaatcatccgctaccttcacgcca
15000 atggcgcctcaatattcttatctgcctcttcctacacatcgggcgaggc
15050 ctatattacggatcatttctctactcagaaacctgaaacatcggcattat
15100 cctcctgcttgcaactatagcaacagccttcataggctatgtcctcccgt
15150 gaggccaaatatcattctgaggggccacagtaattacaaacttactatcc
15200 gccatcccatacattgggacagacctagttcaatgaatctgaggaggcta
15250 ctcagtagacagtcccacccctcacacgattctttacctttcacttcatct
15300 tgcccttcattattgcagccctagcaacactccacctcctattcttgcac
15350 gaaacgggatcaaacaacccccctaggaatcacctcccattccgataaaat
15400 caccttccacccttactacacaatcaaagacgccctcggcttacttctct
15450 tccttctctccttaatgacattaacactattctcaccagacctcctaggc
15500 gacccagacaattataccctagccaacccccttaaacacccctccccacat
15550 caagcccgaatgatatttcctattcgcctacacaattctccgatccgtcc
15600 ctaacaaactaggaggcgtccttgccctattactatccatcctcatccta
15650 gcaataatccccatcctccatatatccaaacaacaaagcataatatttcg
15700 cccactaagccaatcactttattgactcctagccgcagacctcctcattc
15750 taacctgaatcggaggacaaccagtaagctaccctttaccatcattgga
15800 caagtagcatccgtactatacttcacaacaatcctaatcctaataccaac
15850 tatctccctaattgaaaacaaaatactcaaatgggcctgtccttgtagta
15900 taaactaatacaccagtcttgtaaaccggagatgaaaaccttttccaag
15950 gacaaatcagagaaaaagtctttaactccaccattagcacccaaagctaa
14766 (2)
15301 (3)
15607 (2)
14470 (2)
14798 (2)
15043 (2)
15236 (2)
15244 (2)
15784 (2)
15924 (3)
21F
20R
22F
21R
23F
22R


tRNA-Pro
(15957-16023)
Control region
16024-577
16000 gattctaatttaaactattctctgttctttcatggggaagcagatttggg
16050 taccacccaagtattgactcacccatcaacaaccgctatgtatttcgtac
16100 attactgccagccaccatgaatattgtacggtaccataaatacttgacca
16150 cctgtagtacataaaaacccaatccacatcaaaaccccctccccatgctt
16200 acaagcaagtacagcaatcaaccctcaactatcacacatcaactgcaact
16250 ccaaagccacccctcacccactaggataccaacaaacctacccacccttа
16300 acagtacatagtacataaagccatttaccgtacatagcacattacagtca
16350 aatcccttctcgtccccatggatgacccccctcagataggggtcccttga
16400 ccaccatcctccgtgaaatcaatatcccgcacaagagtgctactctcctc 24F
16450 gctccgggcccataacacttgggggtagctaaagtgaactgtatccgaca
16500 tctggttcctacttcagggtcataaagcctaaatagcccacacgttcccc
16550 ttaaataagacatcacgatg
1 gatcacaggtctatcaccctattaaccactcacgggagctctccatgca 23R
50 tttggtattttcgtctgggggggtatgcacgcgatagcattgcgagacgct
100 ggagccggagcaccctatgtcgcagtatctgtctttgattcctgcctcat
150 cctattatttatcgcacctacgttcaatattacaggcgaacatacttact
200 aaagtgtgttaattaattaatgcttgtaggacataataataacaattgaa
250 tgtctgcacagccactttccacacagacatcataacaaaaaatttccacc
300 aaaccccccctcccccgcttctggccacagcacttaaacacatctctgcc
350 aaaccccaaaaacaaagaaccctaacaccagcctaaccagatttcaaatt
400 ttatcttttggcggtatgcacttttaacagtcaccccccaactaacacat
450 tattttcccctcccactcccatactactaatctcatcaatacaacccccg
500 cccatcctacccagcacacacacaccgctgctaaccccataccccgaacc
550 aaccaaaccccaaagacacccccacagtttatgtagcttacctcctcaa 1F

Appendix C. Methodology details

C2.1 Molecular methodology details 158
C5.1 PAUP* commands 75-taxon analysis 161
C5.2 MMS parameters 75-taxon analysis 162
C5.3 C++ code for random selection of 15 taxa from globalhapscoded.nex 163
C5.4 PAUP* commands for coding vs. HVR-I analysis 164

C2.1.1 PCR reagents and conditions

The mtDNA samples for complete sequencing were amplified in two steps: the first a long product polymerase chain reaction (long PCR) which allowed the complete mt genome to be amplified in two overlapping fragments, named Fragment 1 and Fragment 2. The majority of the long PCR reactions were carried out using the Expand Long Template PCR System (Roche Applied Science, Mannheim, Germany) following the manufacturer's instructions. The reactions had a total volume of 50µL made up in two separate master mixes which were thoroughly mixed before beginning thermocycling. A second enzyme mix designed to amplify long products the Triplemaster® PCR system (Eppendorf, Germany) was also used for some reactions and showed a similarly high rate of successful amplification.

	Reagent	µl
Long PCR mix 1:		
	2mM dNTPs	12.5
	Primer 1 (10 pmol/µL)	2
	Primer 2 (10 pmol/µL)	2
	DNA template	2-10
	H_2O	to 25
Long PCR mix 2:		
	10x Buffer 3	5.0
	Taq Expand (5Uµ/L, Roche)	0.75
	H_2O	19.5

The primer combinations used to amplify the entire mt genomes in two overlapping products were 1F and 11R to amplify 'Fragment 1' and 11F and 1R for 'Fragment 2', from the set of 24 primer pairs (Table C2.1) described by Rieder et al (1998) These combinations generated products of length 7517bp and 10 861bp respectively, according to the rCRS. Other primer combinations were tried, but resulted in multiple products or no amplification.

In the second amplification step 12 smaller segments of ~2kb were amplified in 20µL volumes from the large fragments. In some cases it was possible to sequence directly from the long fragments; however most of the sequence data was generated from the segment products. These were named A-L, and the primer sets used for

Table C2.1. Primers used for mt genome amplification (Rieder et al 1998)

Primer	Sequence (5' to 3': L strand)
1F	CTCCTCAAAGCAATACACTG
1R	TGCTAAATCCACCTTCGACC
2F	CGATCAACCTCACCACCTCT
2R	TGGACAACCAGCTATCACCA
3F	GGACTAACCCCTATACCTTCTGC
3R	GGCAGGTCAATTTCACTGGT
4F	AAATCTTACCCCGCCTGTTT
4R	AGGAATGCCATTGCGATTAG
5F	TACTTCACAAAGCGCCTTCC
5R	ATGAAGAATAGGGCGAAGGG
6F	TGGCTCCTTTAACCTCTCCA
6R	AAGGATTATGGATGCGGTTG
7F	ACTAATTAATCCCCTGGCCC
7R	CCTGGGGTGGGTTTTGTATG
8F	CTAACCGGCTTTTTGCCC
8R	ACCTAGAAGGTTGCCTGGCT
9F	GAGGCCTAACCCCTGTCTTT
9R	ATTCCGAAGCCTGGTAGGAT
10F	CTCTTCGTCTGATCCGTCCT
10R	AGCGAAGGCTTCTCAAATCA
11F	ACGCCAAAATCCATTTCACT
11R	CGGGAATTGCATCTGTTTTT
12F	ACGAGTACACCGACTACGGC
12R	TGGGTGGTTGGTGTAAATGA
13F	TTTCCCCCTCTATTGATCCC
13R	GTGGCCTTGGTATGTGCTTT
14F	CCCACCAATCACATGCCTAT
14R	TGTAGCCGTTGAGTTGTGGT
15F	TCTCCATCTATTGATGAGGGTCT
15R	AATTAGGCTGTGGGTGGTTG
16F	GCCATACTAGTCTTTGCCGC
16R	TTGAGAATGAGTGTGAGGCG
17F	TCACTCTCACTGCCCAAGAA
17R	GGAGAATGGGGGATAGGTGT
18F	TATCACTCTCCTACTTACAG
18R	AGAAGGTTATAATTCCTACG
19F	AAACAACCCAGCTCTCCCTAA
19R	TCGATGATGTGGTCTTTGGA
20F	ACATCTGTACCCACGCCTTC
20R	AGAGGGGTCAGGGTTCATTC
21F	GCATAATTAAACTTTACTTC
21R	AGAATATTGAGGCGCCATTG
22F	TGAAACTTCGGCTCACTCCT
22R	AGCTTTGGGTGCTAATGGTG
23F	TCATTGGACAAGTAGCATCC
23R	GAGTGGTTAATAGGGTGATAG
24F	CACCATTCTCCGTGAAATCA
24R	AGGCTAAGCGTTTTGAGCTG

these are listed in Table C2.2.

Segment PCR reagents:

Reagent	µl
10x Buffer	2.0
2mM dNTPs	2.5
Betaine	4
Primer 1 (10 pmol/µL)	1
Primer 2 (10 pmol/µL)	1
DNA	1
Taq (5U/µL)	0.2
H_2O	8.3

The PCR products were cleaned up using filtration plates (Millipore Montage Plates, Billerica, MA, USA) either by centrifuge or vacuum, and products were resuspended in water. The product yield was estimated by visualisation on an agarose gel stained with ethidium bromide. Examples of long fragment and segment amplifications are shown in Figure C2.1.

PCR conditions:

Long PCR	
Initial denaturation	93°C 2 minutes
10 cycles	93°C 30 seconds
	55°C 30 seconds
	68°C 8 minutes
25 cycles	93°C 30 seconds
	55°C 30 seconds
	68°C 8 minutes + 20 sec/cycle
Final elongation	68°C 11 minutes

Segment PCR	
Initial denaturation	94°C 2 minutes
34 cycles	94°C 30 seconds
	55°C 30 seconds
	72°C 3 minutes
Final elongation	72°C 5 minutes

Table C2.2. Primer combinations used for segment amplifications

PCR Segment	PCR primers	Length (kb)
A	1F, 3R	2
B	3F, 5R	2.1
C	5F, 7R	2.2
D	7F, 9R	2.1
E	9F, 11R	2.2
F	11F, 13R	2.2
G	13F, 15R	2.2
H	15F, 17R	2
I	17F, 19R	2.2
J	19F, 21R	2.4
K	21F, 22R	1.9
L	22F, 24R	2.5

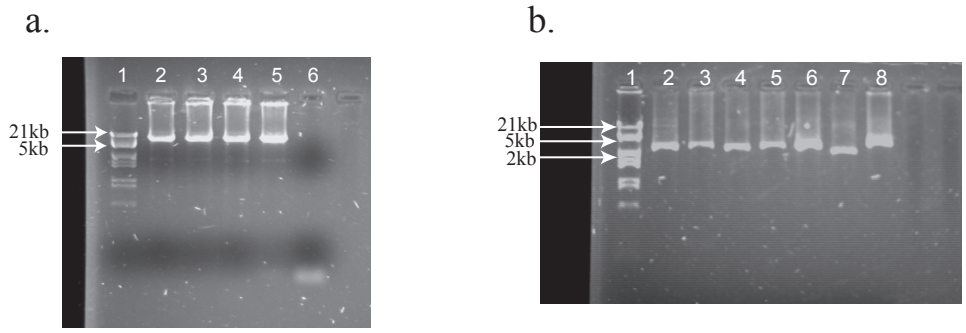


Figure C2.1: Visualisation of PCR products
a) Long Fragment 1 amplification products from four samples (laboratory notebook 1, PCR 12, p31). Lanes left to right: 1. λ HindIII/EcoR1 molecular weight marker, 2. MJ86, 3. AMI15, 4. MJ22, 5. MO304, 6. H_2O control.
b) Seven segment amplifications from sample MJ86 (laboratory notebook 1, PCR 19 p47). Lanes left to right: 1. λ Hind III/EcoR1 molecular weight marker, 2. Segment F, 3. Segment G, 4. Segment H, 5. Segment I, 6. Segment J, 7. Segment K, 8. Segment L.

C2.1.2 Sequencing methods

Sequencing was carried out in both forward and reverse directions, using the complete set of primers. The BDT (BigDye Terminator chemistry, Applied Biosystems, Foster City, CA, USA) version 3.1 protocol was followed, using 1/8x reactions:

Reagent	µL
BDT	1
5 x Seq Buffer	3.5
Primer (1 µM)	3.2
Template	1-8
H_2O	to 20

Sequencing reaction conditions:	25 cycles	96ºC 10 seconds
		50ºC 10 seconds
		60ºC 60 seconds
	Hold at 15ºC	

The sequence reaction products were cleaned by centrifugation through a sephadex column, and analysed on an ABI3730 Genetic Analyzer (Applied Biosystems Inc.) at the Allan Wilson Centre Genome Service (Palmerston North, New Zealand). The set of samples from Auckland which were sequenced across the control region only were analysed by Canterbury Sequencing in the School of Biological Sciences, University of Canterbury on an ABI3100 Genetic Analyzer (Applied Biosystems Inc.).

Sequence electropherograms were edited and mt genomes assembled using Sequencher™ (Version 4.2.2, Gene Codes Corporation).

C5.1 PAUP* commands for heuristic tree searches, 75-taxa analysis

```
BEGIN PAUP;
LOG FILE=1_50.log;
SET STATUS=NO;
SET NOTIIFYBEEP=NO:
SET CRITERION=PARSIMONY;
SET INCREASE=AUTO;
SET AUTOCLOSE=YES;
EXCLUDE 1837-2012;
EXCLUDE 1-142;
EXCLUDE CONSTANT;
EXCLUDE UNINF;
EXPORT FILE=1_50MMS.phy FORMAT=PHYLIP;
HSEARCH;
SAVETREES FILE=1_50PHY.tre FORMAT=PHYLIP;
INCLUDE ALL;
PSCORES ALL /SCOREFILE=1_50pscores.txt single=all displayout=no;
PSCORES [ALL] /SINGLE=VAR TOTAL=NO DISPLAYOUT=YES;
CLEARTREES [NOWARN=YES];
;
END;
```

Note: Separate outfiles (log file, PHYLIP data set, trees and parsimony scores) were generated for each of the 50 data sets in this analysis, requiring the names of the files to be changed for each PAUP* block (for example 1_50.log, 2_50.log, 3_50.log). These were prepared using a combination of BBEdit Lite (version 6.1.2 Bare Bones Software, MA, USA) and Excel®· functions. Appending the results from successive data sets to the first using the APPEND=YES command, as in the analysis of coding vs. HVR-I below, would be a more straightforward approach.

C5.2 MMS parameters for six successive rounds used on 200 data sets of 75-taxon analysis

1. 100 iterations, 50 nochanges, generatepart=likelygroup
2. 100 iterations, 75 nochanges, generatepart=likelygroup
3. 150 iterations nochanges=200, generatepart=likelygroup
4. 200 iterations, 250 nochanges, generatepart=likelygroup
5. 50 iterations, 300 nochanges, generatepart=likelygroup
6. 100 iterations, 200 nochanges, generatepart=likelygroup

C5.3 C++ program to select random n=15 data sets from globalhapcoded.nex (D.Bryant, University of Auckland).

```
#include <iostream>

#include <ext/algorithm>
#include<vector>

using namespace std;

int main (int argc, char * const argv[]) {

	const int ngroups = 18; //Number of groups in total
	const int nsampled = 3; //Number of groups to sample
	const int samplesPerGroup = 5; //Number of samples to select for each group
	const int nreplicates = 9000; //Number of replicates to perform
	int hapSizes[ngroups] = {30,47,19,17,261,63,88,51,16,56,45,28,254,121,17,55,164,32}; //Number taxa in each groups

	//ASSUMPTION: taxa 1..30 in group 1, 31..77 in group 2, etc.

	//Form vectors of the taxa in each group.
	int taxon = 1;
	vector< vector<int> > grouptaxa(ngroups);

	for(int g=0;g<ngroups;g++) {
		grouptaxa[g].clear();
		for(int i=0;i<hapSizes[g];i++) {
			grouptaxa[g].push_back(taxon);
			taxon++;
		}
	}

	//Another vector of the actual group choices

	vector<int> groups (ngroups);
	for(int g=0;g<ngroups;g++)
		groups[g]=g;


	//Now do the selections

	for(int i = 0;i<nreplicates; i++) {
		vector<int> groupsSampled(nsampled);
		random_sample(groups.begin(),groups.end(), groupsSampled.begin(),groupsSampled.end());
```

```
		vector<int> sample(nsampled*samplesPerGroup);
		vector<int>::iterator s = sample.begin();

		//Now choose taxa for each group/
		for(int j=0;j<groupsSampled.size();j++) {
			vector<int> thegroup = grouptaxa[groupsSampled[j]];
			//cout<<"Group "<<groupsSampled[j]+1<<" goes from "<<thegroup.front()<<" to
"<<thegroup.back()<<endl;


			random_sample(thegroup.begin(), thegroup.end(), s, s+samplesPerGroup);

		//	cout<<"Sample is ";
		//	for(vector<int>::iterator p = s; p!=s+samplesPerGroup; p++)
		//		cout<<(*p)<<" ";
		//	cout<<endl;

			s+=samplesPerGroup;
		}

		cout<<"\n\n[Replicate "<<i+1<<".\t Groups";
		for(int j=0;j<nsampled;j++)
			cout<<" "<<groupsSampled[j]+1;
		cout<<"]"<<endl;

		cout<<"REPLACEMENT1"<<endl;
		cout<<endl;

		cout<<"\tUNDELETE";
		for(int j=0;j<sample.size();j++) {
			cout<<" "<<sample[j];
		}
		cout<<" / ONLY CLEARTREES;"<<endl;

		cout<<"\nREPLACEMENT2"<<endl;

	}

  return 0;
}
```

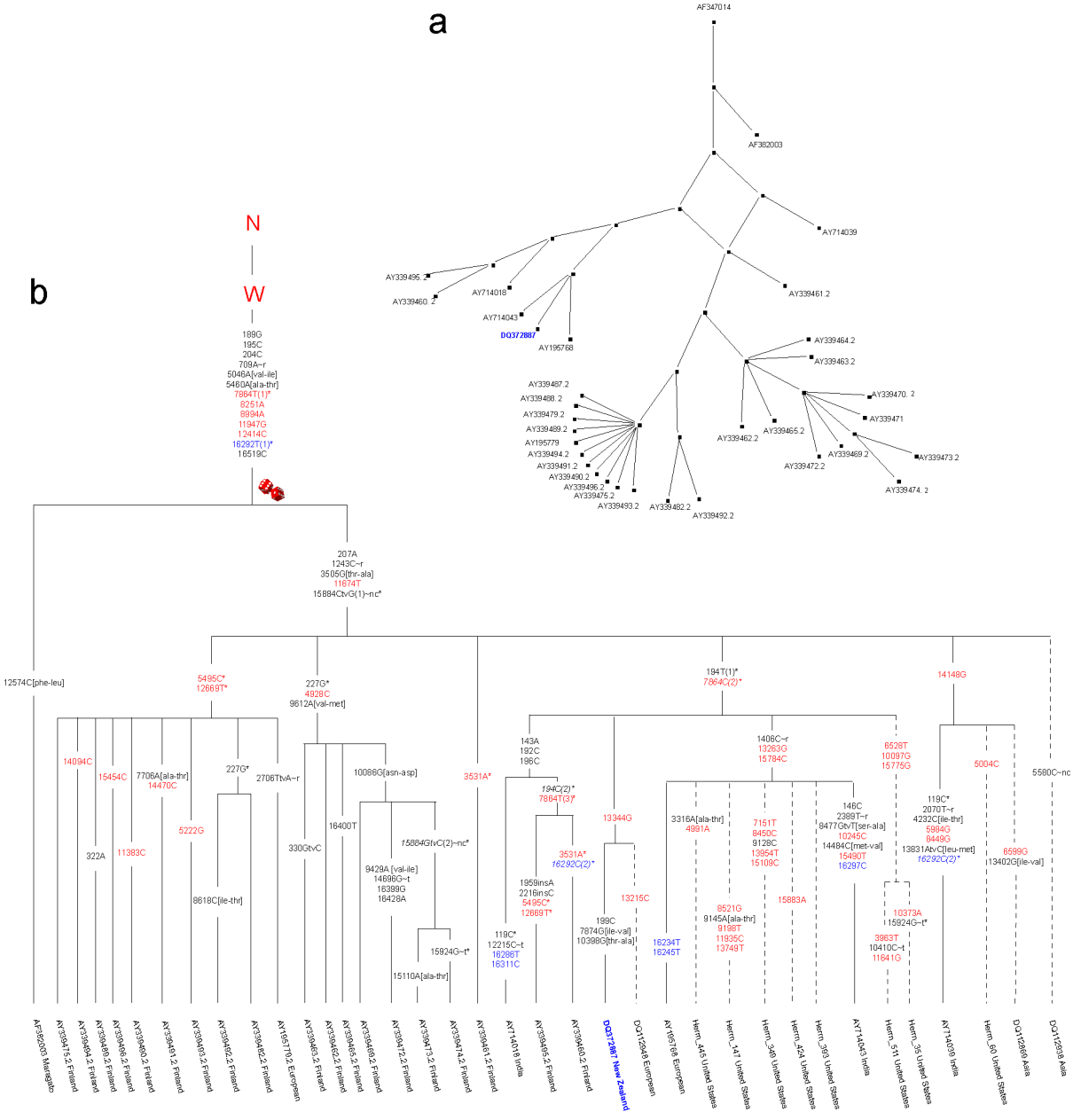
C5.4 PAUP* commands for heuristic tree searches, coding vs. HVR-I analysis

(Note: the 'UNDELETE' command lists the 15 random taxa to be included in the analysis; taxa 1-15 are listed here for example purposes.

```
BEGIN PAUP;
SET STATUS=NO;
SET NOTIFYBEEP=NO:
LOG FILE=globaltrip.log  APPEND=YES;
UNDELETE 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 / ONLY
WTSET * UNTITLED  = 1: 1-16711;
SET CRITERION=PARSIMONY;
SET INCREASE=AUTO;
SET AUTOCLOSE=YES;
EXCLUDE 16142-16711;
EXCLUDE 1-617;
EXCLUDE CONSTANT;
EXCLUDE UNINF;
HSEARCH;
EXCLUDE ALL;
INCLUDE 16694-16711;
PSCORES 1/SCOREFILE=codingpscores.txt APPEND=YES single=all displayout=no;
CLEARTREES [NOWARN=YES];
INCLUDE ALL;
EXCLUDE 1-16141;
EXCLUDE 16506-16711;
EXCLUDE CONSTANT;
EXCLUDE UNINF;
HSEARCH;
EXCLUDE ALL;
INCLUDE 16694-16711;
PSCORES 1/SCOREFILE=HVRIpscores.txt APPEND=YES single=all displayout=no;
CLEARTREES [NOWARN=YES];
INCLUDE ALL;
WTSET * HVRIone = 1: 16142-16168 16170-16203 16205-16209 16212-16228 16230-16231 16233-16243 16245-16246 16248-16262
16264-16265 16267-16290 16292-16300 16304-16305 16309-16329 16331-16333 16337 16339-16343 16346-16354 16356-16369
16371-16376 16378-16381 16383-16385 16389-16391 16393-16395 16397-16399 16401-16407 16409-16412 16417-16420 16422-
16426 16428-16431 16433 16435-16441 16444-16447 16449 16451-16477 16479-16484 16486-16490 16492-16506, 0.5: 16169 16210
16229 16232 16263 16266 16303 16307 16334-16336 16338 16345 16370 16377 16382 16386-16388 16392 16396 16414 16443 16448
16450 16478 16491, 0.33: 16204 16244 16301 16330 16355 16408 16413 16421 16427 16432 16442, 0.25: 16211 16306 16415-16416,
0.2: 16291 16344 16400, 0.14: 16302 16485, 0.13: 16434, 0.1: 16247, 0.08: 16308;
EXCLUDE 1-16141;
EXCLUDE 16506-16711;
EXCLUDE CONSTANT;
EXCLUDE UNINF;
HSEARCH;
EXCLUDE ALL;
INCLUDE 16694-16711;
PSCORES 1/SCOREFILE=HVRIw1pscores.txt APPEND=YES single=all displayout=no;
CLEARTREES [NOWARN=YES];
INCLUDE ALL;
WTSET * HVRItwo = 1: 16142-16168 16170-16203 16205-16209 16212-16228 16230-16231 16233-16243 16245-16246 16248-16262
16264-16265 16267-16290 16292-16300 16304-16305 16309-16329 16331-16333 16337 16339-16343 16346-16354 16356-16369
16371-16376 16378-16381 16383-16385 16389-16391 16393-16395 16397-16399 16401-16407 16409-16412 16417-16420 16422-
16426 16428-16431 16433 16435-16441 16444-16447 16449 16451-16477 16479-16484 16486-16490 16492-16506, 0.9: 16169 16210
16229 16232 16263 16266 16303 16307 16334-16336 16338 16345 16370 16377 16382 16386-16388 16392 16396 16414 16443 16448
16450 16478 16491, 0.8: 16204 16244 16301 16330 16355 16408 16413 16421 16427 16432 16442, 0.7: 16211 16306 16415-16416,
0.6: 16291 16344 16400, 0.4: 16302 16485, 0.3: 16434, 0.1: 16247, 0.00: 16308;
EXCLUDE 1-16141;
EXCLUDE 16506-16711;
EXCLUDE CONSTANT;
EXCLUDE UNINF;
HSEARCH;
EXCLUDE ALL;
INCLUDE 16694-16711;
PSCORES 1/SCOREFILE=HVRIw2pscores.txt APPEND=YES single=all displayout=no;
CLEARTREES [NOWARN=YES];
END;
```

Appendix D: Supplementary Figures

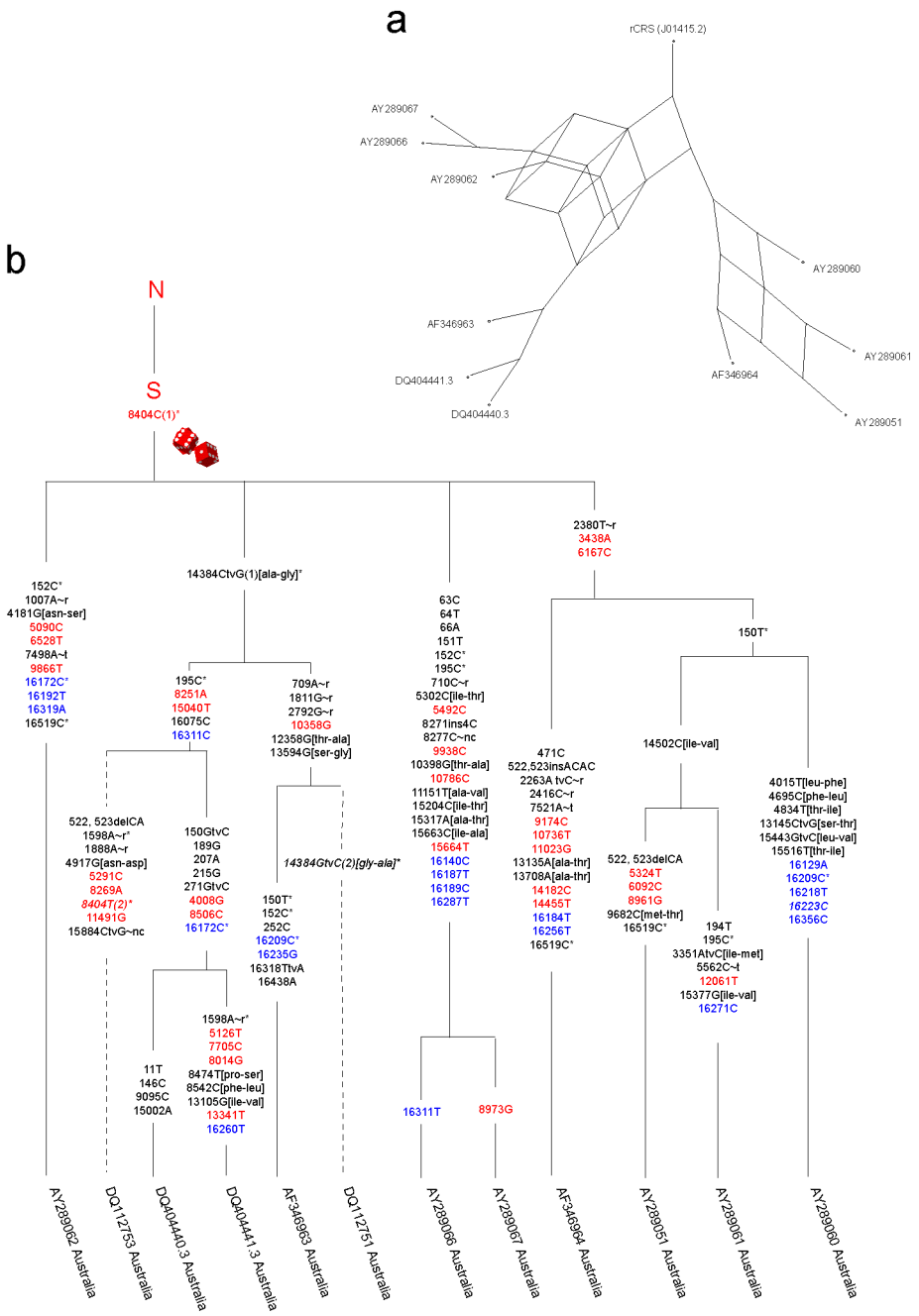
D. Supplementary Figures 165
D3.1 N/W consensus network and labelled phylogeny 166
D3.2 N/S consensus network and labelled phylogeny 167
D3.3 N/R/B4b consensus network and labelled phylogeny 168
D3.4 L1c consensus network and labelled phylogeny 169
D3.5.East Asian skeleton phylogeny macrohaplogroup N (Kong et al 2006) 170
D3.6 East Asian skeleton phylogeny macrohaplogroup M (Kong et al 2006) 171
D4.1 Haplotype tree 172



D3.1 N/W consensus network and labelled phylogeny

a) The consensus network of 15 most parsimonious trees for the 32 N/ W sequences with AF347014 found by heuristic PAUP* (version 4.0b10, Swofford 2003) search, constructed using SplitsTree (version 4.6, Huson and Bryant 2006). The entire mt genome sequence was analysed; there are 24 parsimony informative characters and the parsimony search score of 36 was proved optimal by MMS. The sequence from this study is shown in blue type.

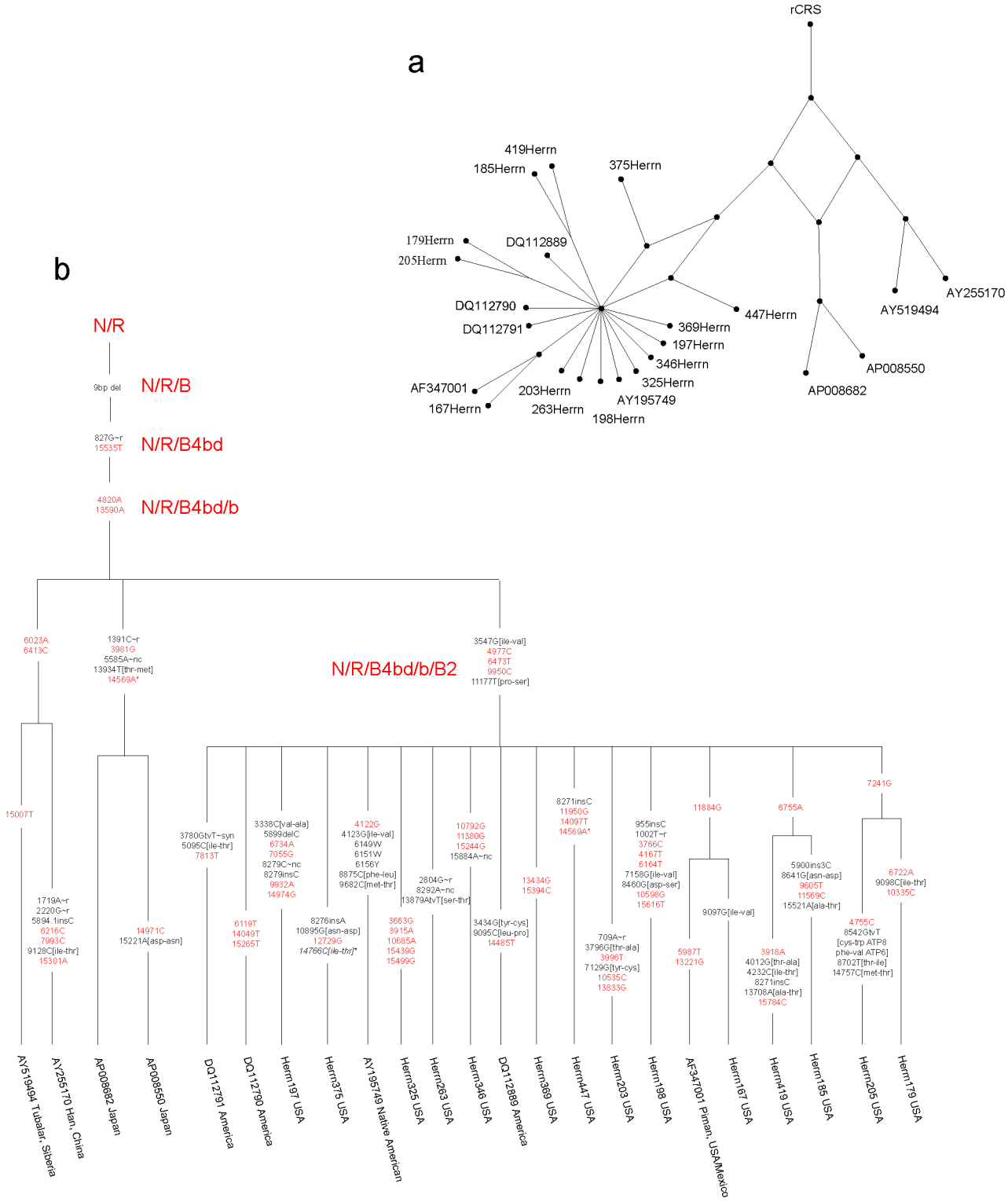
b) A labelled reconstruction of the N/W phylogeny. Dotted branches represent mtDNA coding-region only sequences which were not included in the parsimony analysis. See caption for Figure 3.1 for explanation of abbreviations, colours and codes used. The polymorphisms relative to the rCRS at the N vertex are: 73G, 263G, 750G, 1438G, 2706G, 3106del, 4769G, 7028T, 8860G, 11719A, 12705T, 14766T, 15326G and 16223T. Recurrent or parallel changes occur in this phylogeny at nucleotides 119, 194, 227, 3531, 5495, 7864, 12669, 15884, 15924, and 16292.



Appendix D3.2 N/S haplogroup consensus network and labelled phylogeny

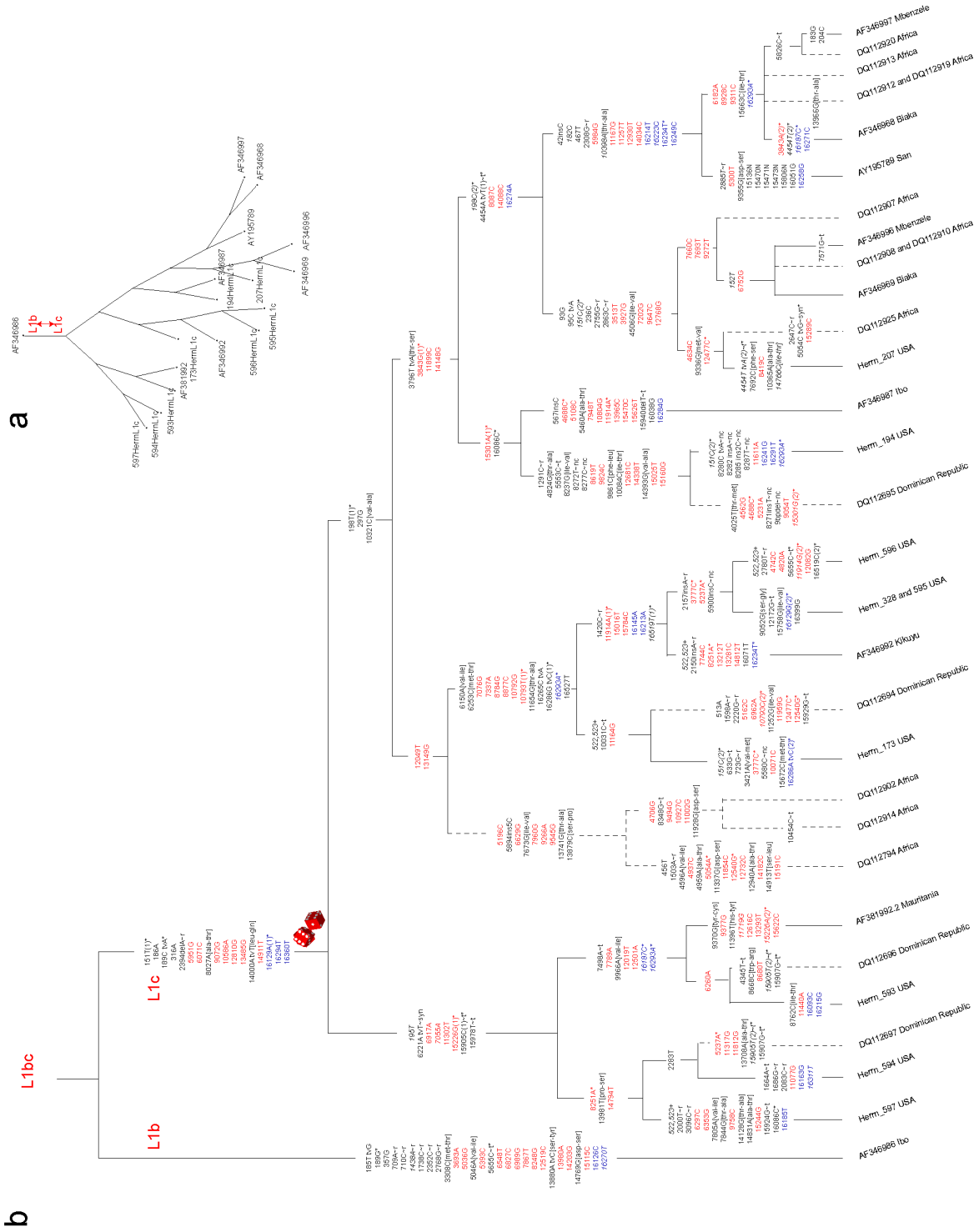
a) The consensus network of 9 most parsimonious trees, from 10 N/S individuals with the rCRS (JO1415.2) found by heuristic PAUP* search, (version 4. 0b10, Swofford 2003), constructed using SplitsTree (version 4.6, Huson and Bryant 2006). The entire mtDNA sequence was analysed: there are 41 parsimony informative characters, and the parsimony score of 52 was proved minimal using MMS.

b) A branch-labelled phylogeny of the N/S haplogroup reconstructed from the consensus network, with Kivisild et al (2005) coding-region sequences added (broken lines). See caption for Figure 3.1 for explanation of abbreviations, colours and codes used. The polymorphisms relative to the rCRS at the N vertex are: 73G, 263G, 750G, 1438G, 2706G, 3106del, 4769G, 7028T, 8860G, 11719A, 12705T, 14766T, 15326G and 16223T. Recurrent or parallel polymorphisms in this phylogeny are at nucleotides 150, 152, 195, 1598, 8404, 14384, 16172, 16209 and 16519. It is possible that the recurrent mutations in sequences DQ112753 and DQ112751 at nt8404 and nt14384 respectively are due to sequencing error. The authors could not confirm the status of the bases at these positions as the original files could not be located (T. Kivisild, email 1/02/07).



Appendix D3.3 N/R/B4b consensus network and branch-labelled phylogeny

a) The consensus network of 3 most parsimonious trees (16 parsimony informative characters; coding region only, parsimony score 18) found by heuristic search (PAUP* version 4.0b10, Swofford 2003) on the B4b data set of 24 sequences including the rCRS. The parsimony score was proved minimal using MMS, and the consensus network constructed using SplitsTree (version 4.6, Huson and Bryant 2006). The sequences labelled xxxHerrn are from Herrnstadt et al (2002), and are available from the mtDB database (http://www.genpat.uu.se/mtDB/, Ingman and Gyllensten 2006).
b) Labelled phylogeny of the B4b sequences reconstructed from the consensus network, with coding region only polymorphisms shown. See caption for Figure 3.1 for explanation of abbreviations, colours and codes used. The coding region polymorphisms relative to the rCRS at the N/R/B4b vertex are: 750G, 827G, 1438G, 2706G, 3106del, 4769G, 4820A, 7028T, 8860G, 11719A, 13590A, 14766T, 15326G, 15535T.
A parallel change occurs in the phylogeny at nt14569 and is marked with an asterisk.



Appendix D3.4 L1c tree and labelled phylogeny

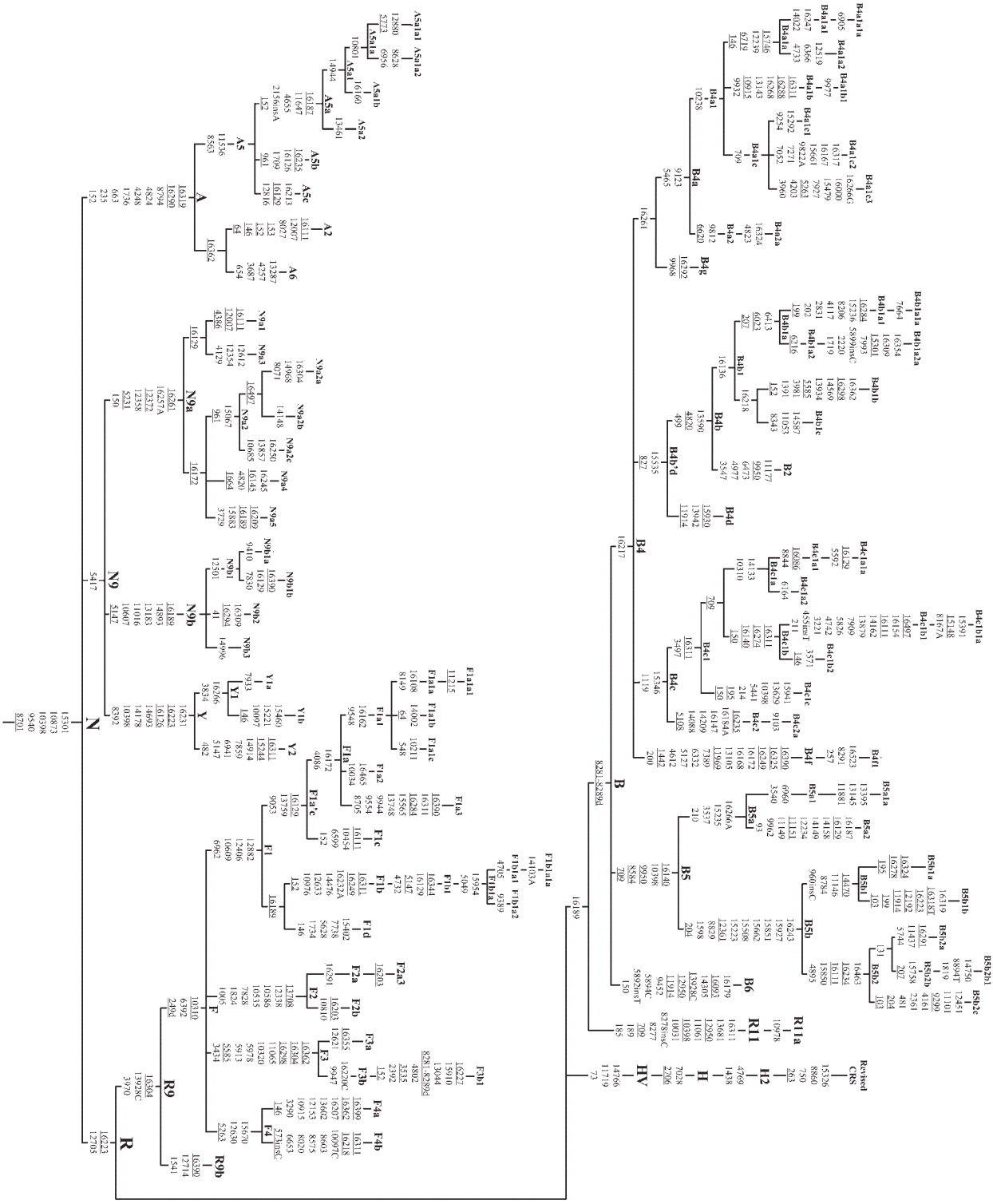
a) The single most parsimonious tree for the 16 L1c and L1b sequences found by heuristic PAUP* (version 4.0b10, Swofford 2003) search, drawn using SplitsTree (version 4.6, Huson and Bryant 2006). The entire mt genome sequence was analysed; there are 94 parsimony informative characters and the parsimony search score of 119 was proved optimal by MMS. Sequences from this study are shown in blue type.

b) A labelled reconstruction of the L1c phylogeny. Dotted branches represent mtDNA coding-region only sequences which were not included in the parsimony analysis. See caption for Figure 3.1 for explanation of abbreviations, colours and codes used. The polymorphisms relative to the rCRS at the L1bc vertex are: 73G, 152C, 182T, 195C, 247A, 263G, 522, 523del, 750G, 769A, 825A, 1018A, 1438G, 2706G, 2758A, 2885C, 3106del, 3594T, 3666A, 4104G, 4769G, 7028T, 7055G, 7146G, 7256T , 7389C, 7521A, 8468T, 8655T, 8701G, 8860G, 9540C, 10398G, 10688A, 10810C, 10873C, 11719A, 12705T, 13105G, 13506T, 13650T, 13789C, 14178C, 14560A, 14766T, 15326G, 16187T, 16189C, 16223T, 16278T, 16293G, 16311C, 16519C. Recurrent or parallel changes occur in this phylogeny at nucleotides 151, 189, 198, 3777, 3843, 4454, 4688, 5054, 5237, 5655, 8251, 10793, 11914, 12477, 12540, 15226, 15301, 15905, 15907, 16086, 16187, 16234.

Appendix D3.5 East Asian skeleton phylogeny macrohaplogroup N

Reproduced from Kong et al ' Updating the East Asian mtDNA phylogeny: a prerequisite for the identification of pathogenic mutations' Human Molecular Genetics, 2006, 15:2080, by permission of Oxford University Press.

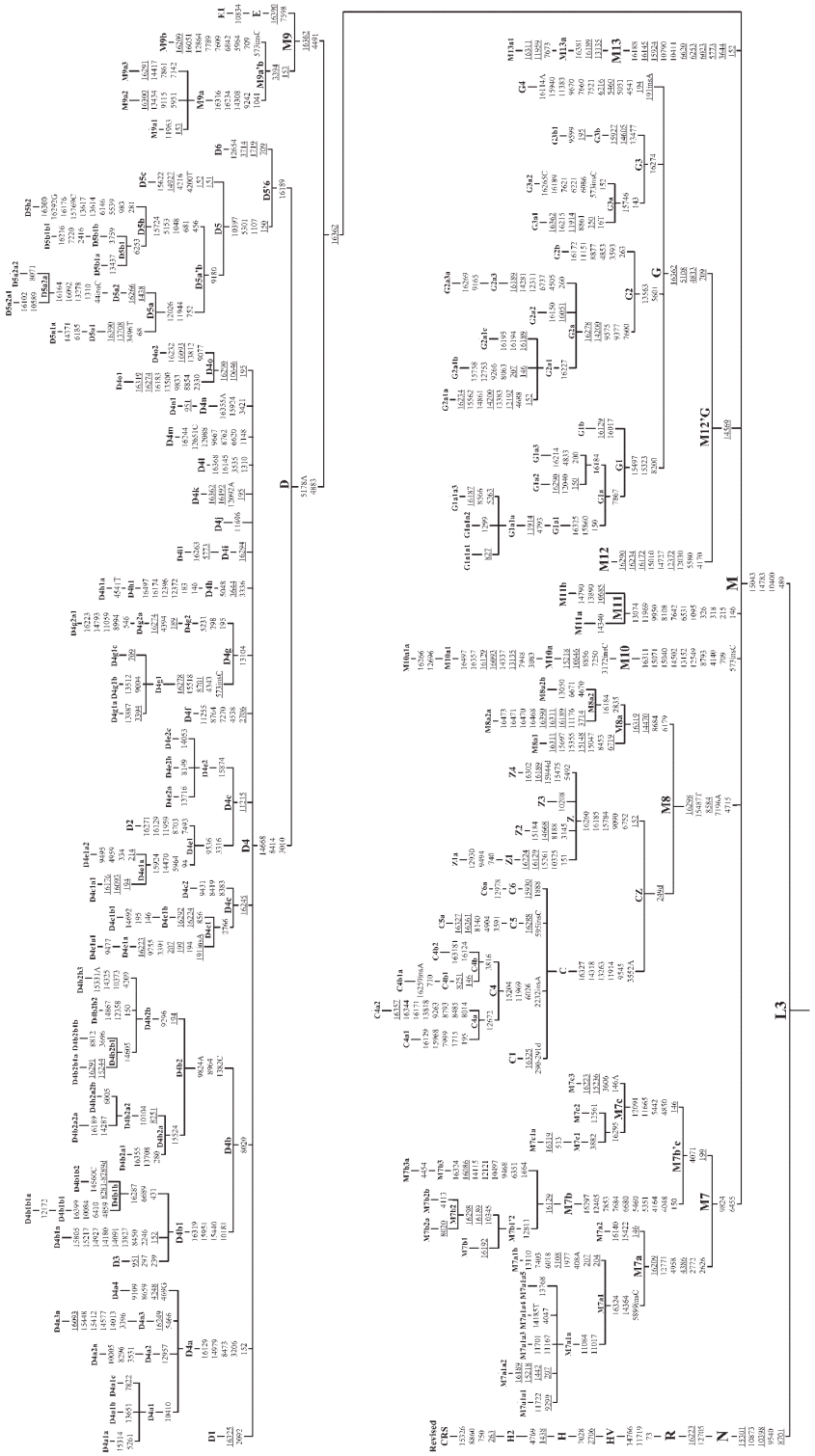
Original figure legend from Kong et al 2006: Updated East Asian mtDNA phylogenetic tree displaying the different subsets of macrohaplogroup N. The A/C stretch length polymorphism in regions 1618016193 and 303315, 522523d and mutation 16519, all known to be hypervariable, were disregarded for tree reconstruction. Suffixes A, C, G and T refer to transversions, 'd' means deletion, and 'ins' indicates an insertion event (the exact number of the inserted nucleotide(s) was disregarded); recurrent mutations are underlined. Because of the lack of information, only the control-region motif of haplogroup G1b was determined.

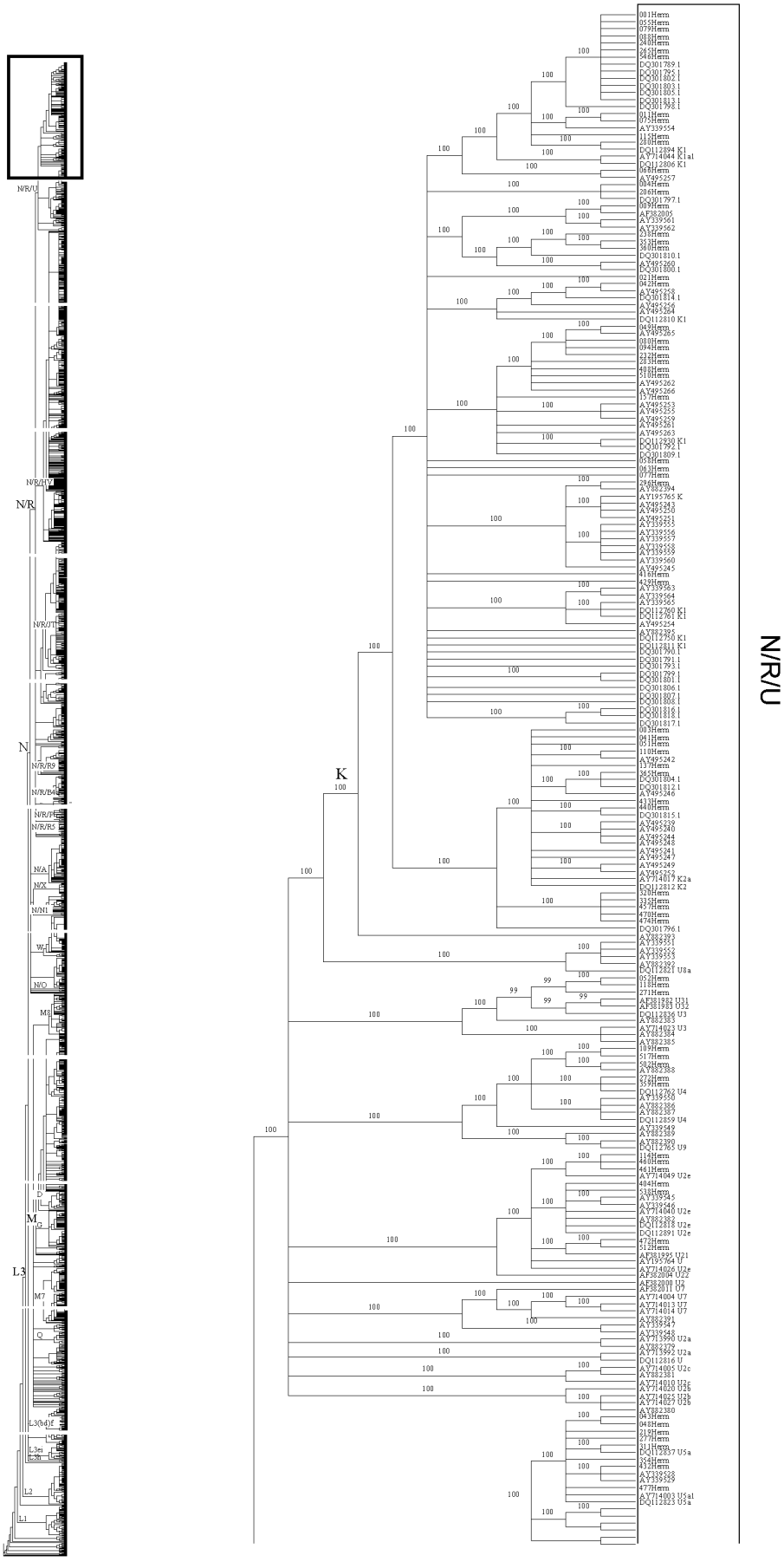


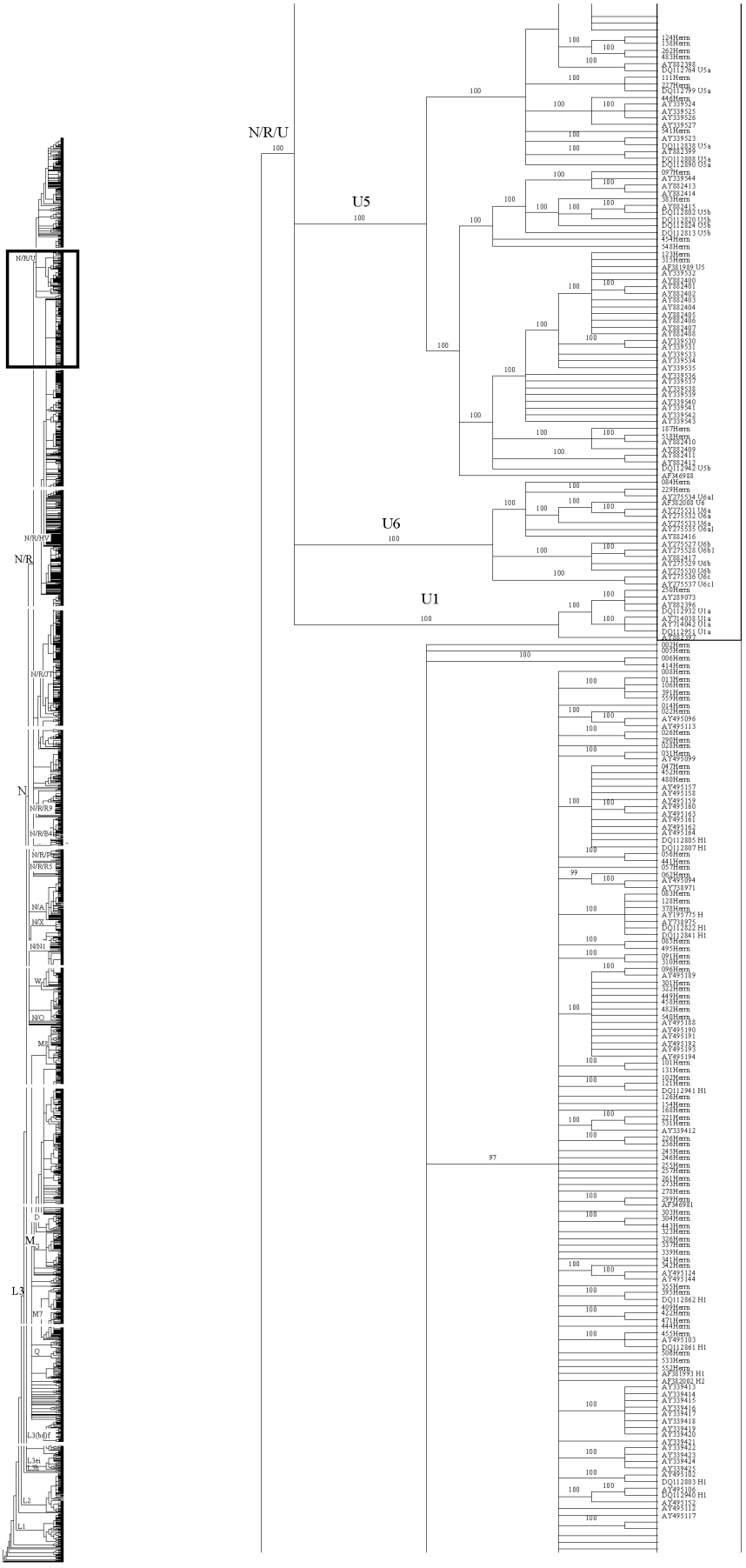
Appendix D3.6 East Asian skeleton phylogeny macrohaplogroup M

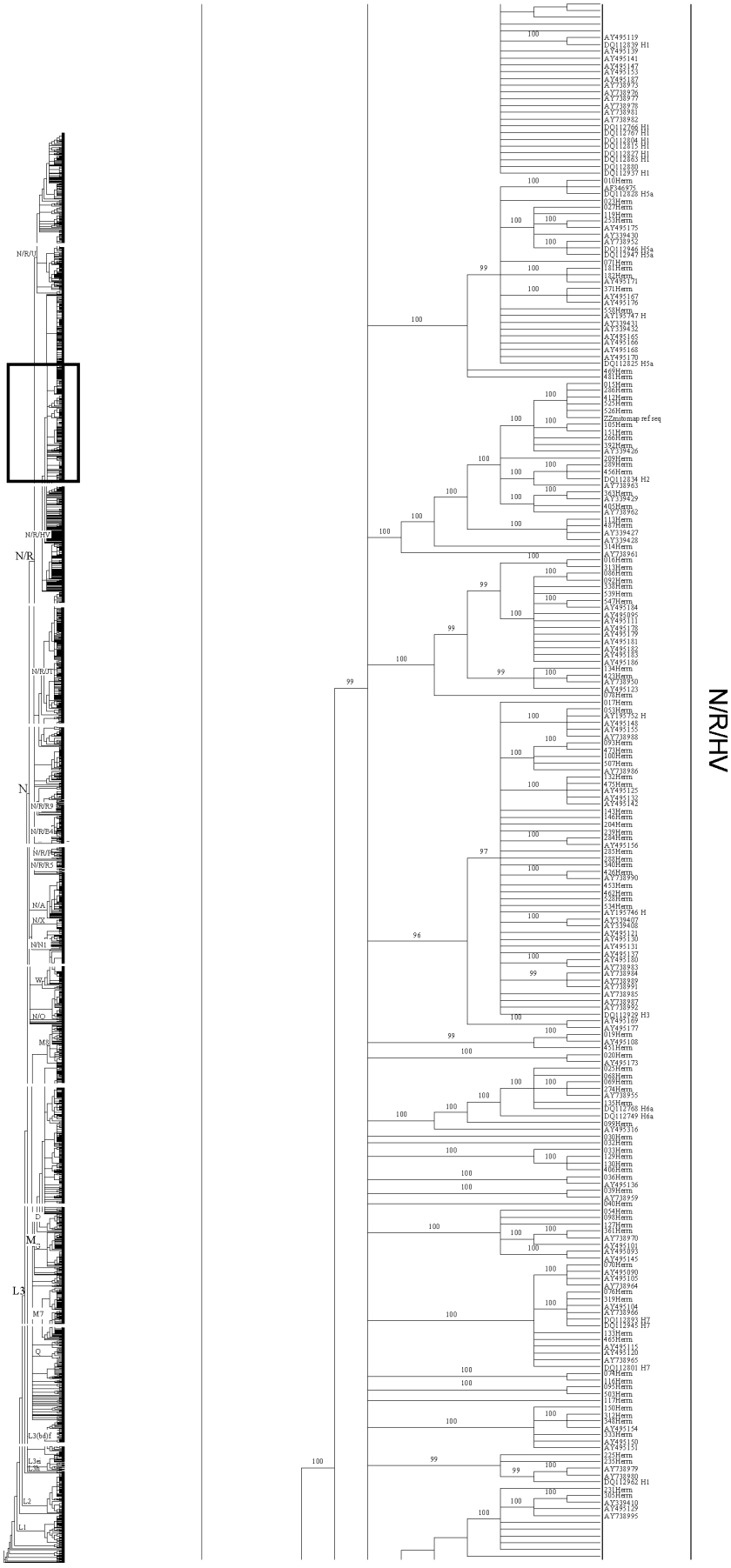
Reproduced from Kong et al ‘ Updating the East Asian mtDNA phylogeny: a prerequisite for the identification of pathogenic mutations’ Human Molecular Genetics, 2006, 15:2078, by permission of Oxford University Press.

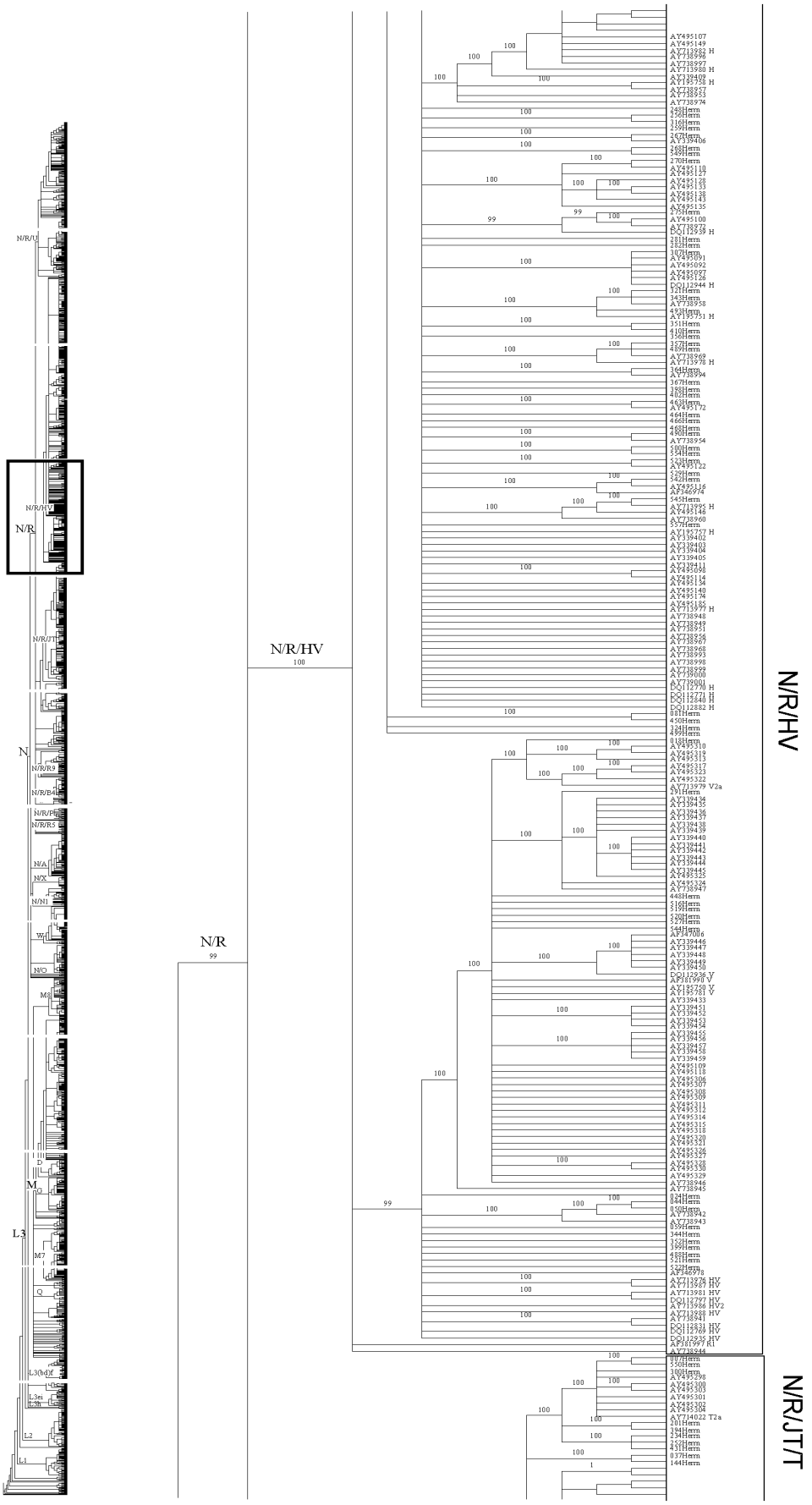
Original figure legend from Kong et al 2006: Updated East Asian mtDNA phylogenetic tree displaying the different subsets of macrohaplogroup M. The A/C stretch length polymorphism in regions 1618016193 and 303315, 522523d and mutation 16519, all known to be hypervariable, were disregarded for tree reconstruction. Suffixes A, C, G and T refer to transversions, ‘d’ means deletion, and ‘ins’ indicates an insertion event (the exact number of the inserted nucleotide(s) was disregarded); recurrent mutations are underlined. Because of the lack of information, only the control-region motif of haplogroup G1b was determined.

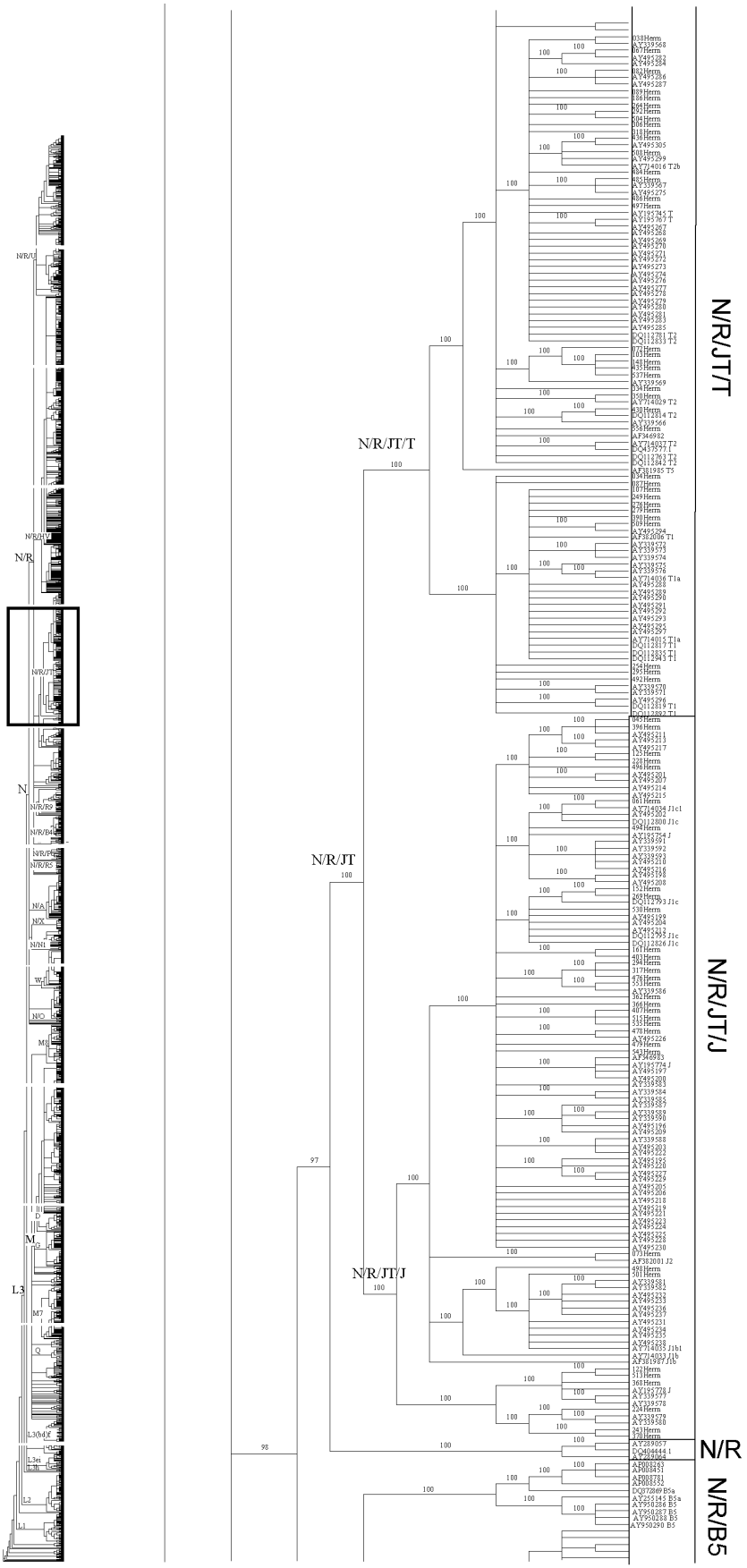


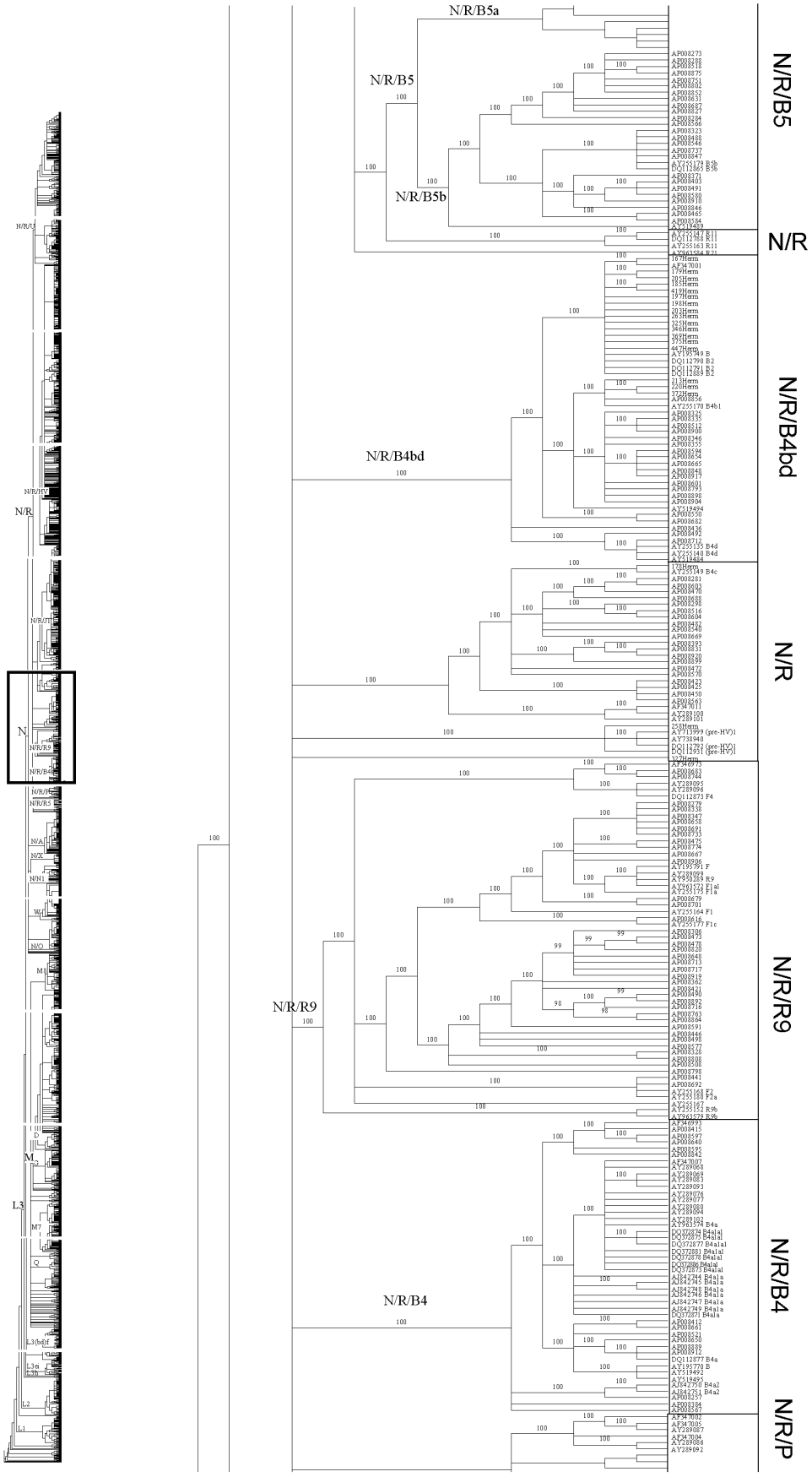

N/R/U
N/R
N
K
M
L3
N/R/U
N/R/U1
N/R/R9
N/R/B4
N/R/R4
N/R/R5
N/A
N/X
N/N1
N/O
M8
D
G
M7
Q
L3(b4)f
L3ej
L3eh
L2
L1

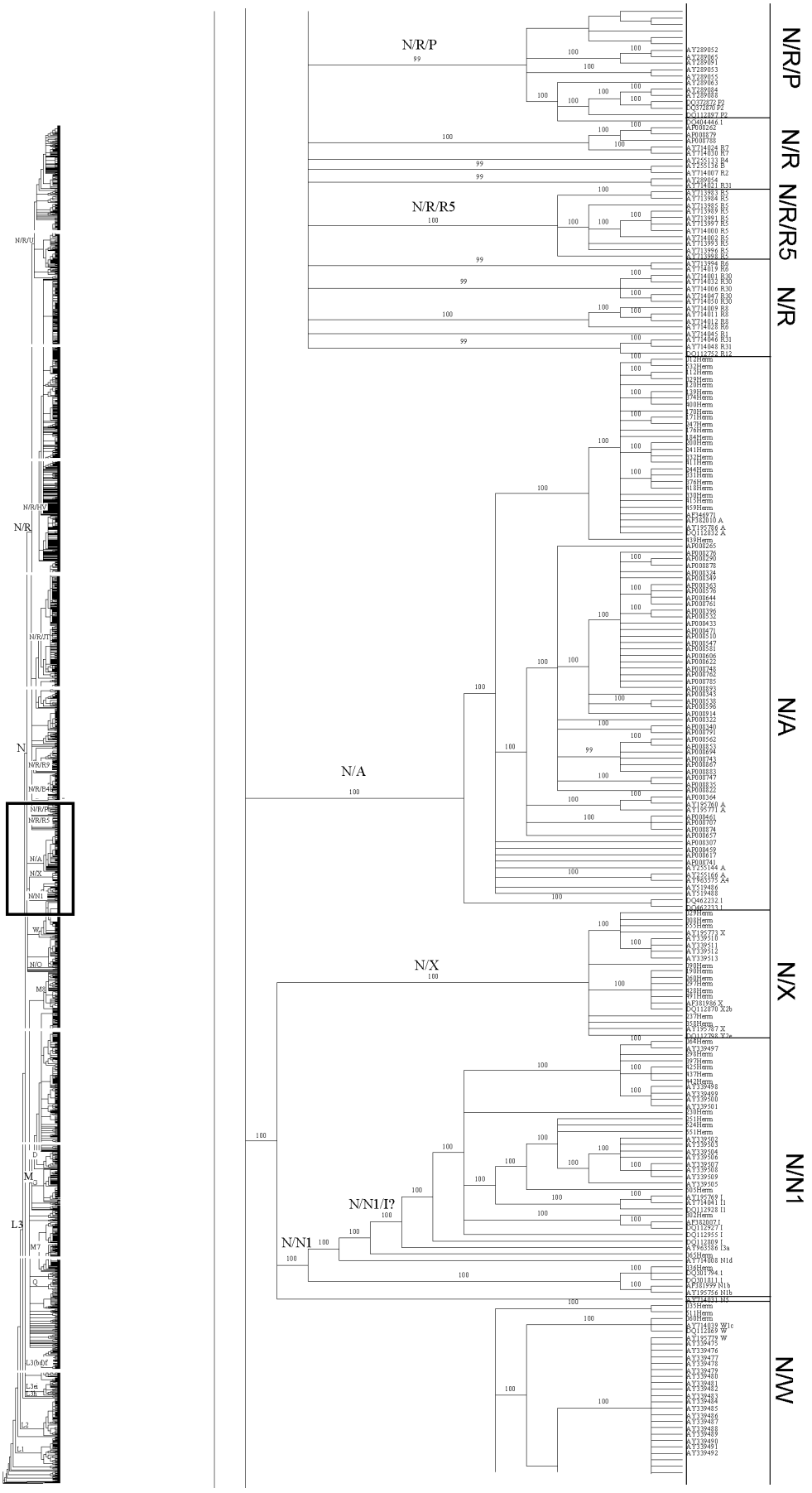

N/R/U
U5
U6
U1
N/R/HV
N/R/U
N/R/HV

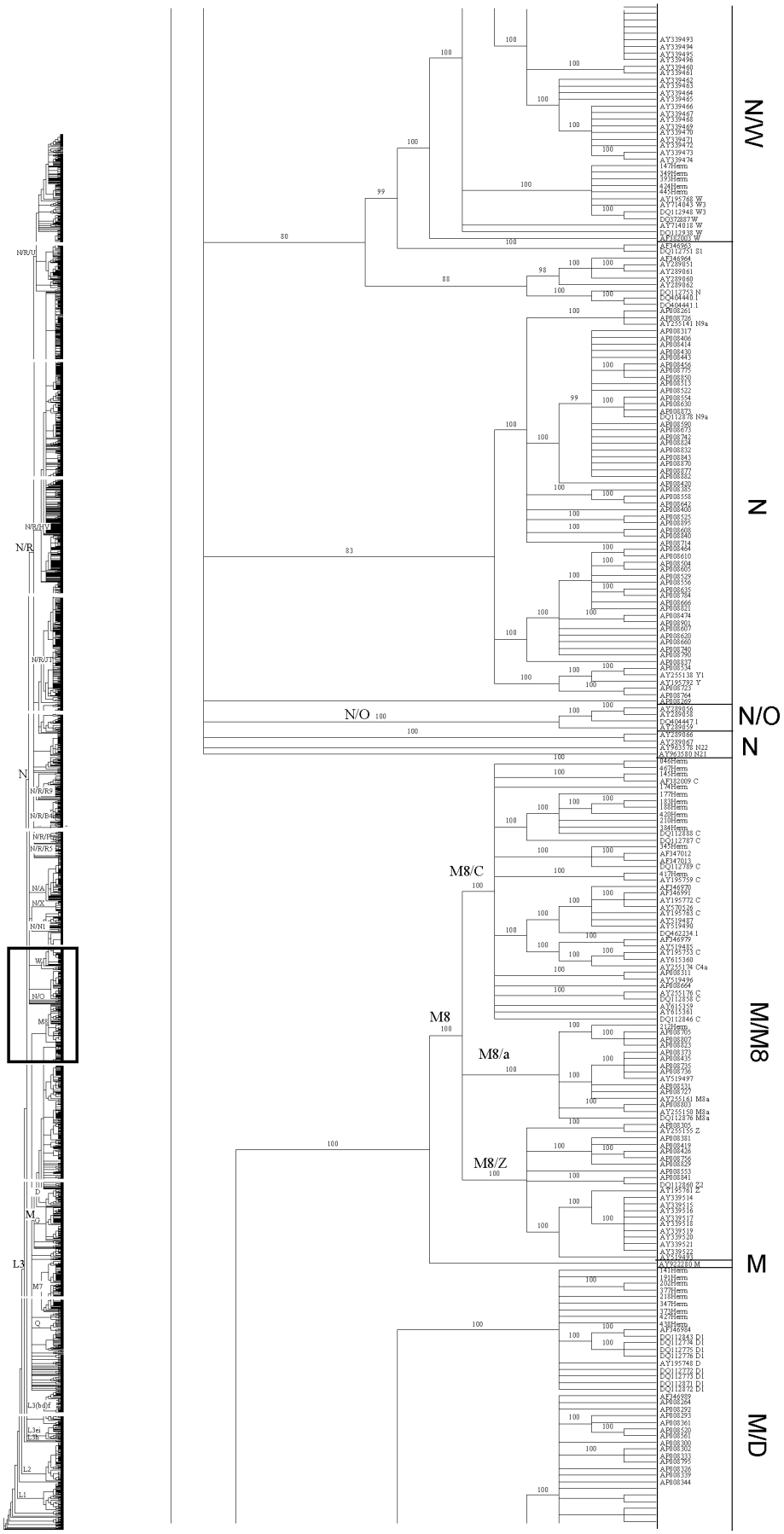

N/R/U
N/R/HV
N/R
N/R/JT
N
N/R/R9
N/R/B4a
L
N/R/R5
N/A
N/X
N/N1
W
N/O
M8
M
n2
L3
M7
Q
L3/b/f
L3e
L3h
L2
L1
N/R/HV

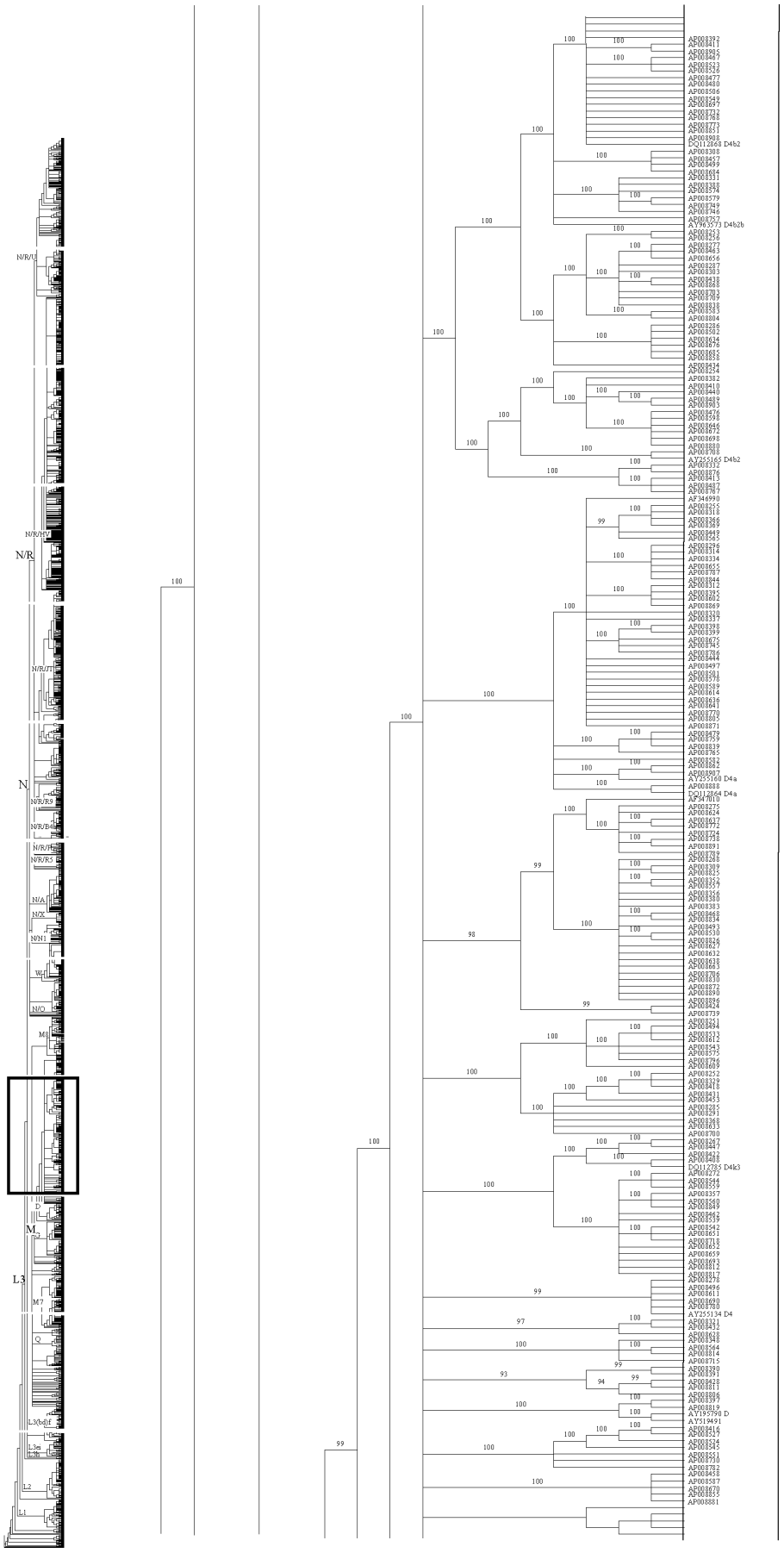

N/R/U
N/R/HV
N/R
N/R/JT
N
M
L3
L2
L1
N/R/HV
N/R
N/R/JT/T

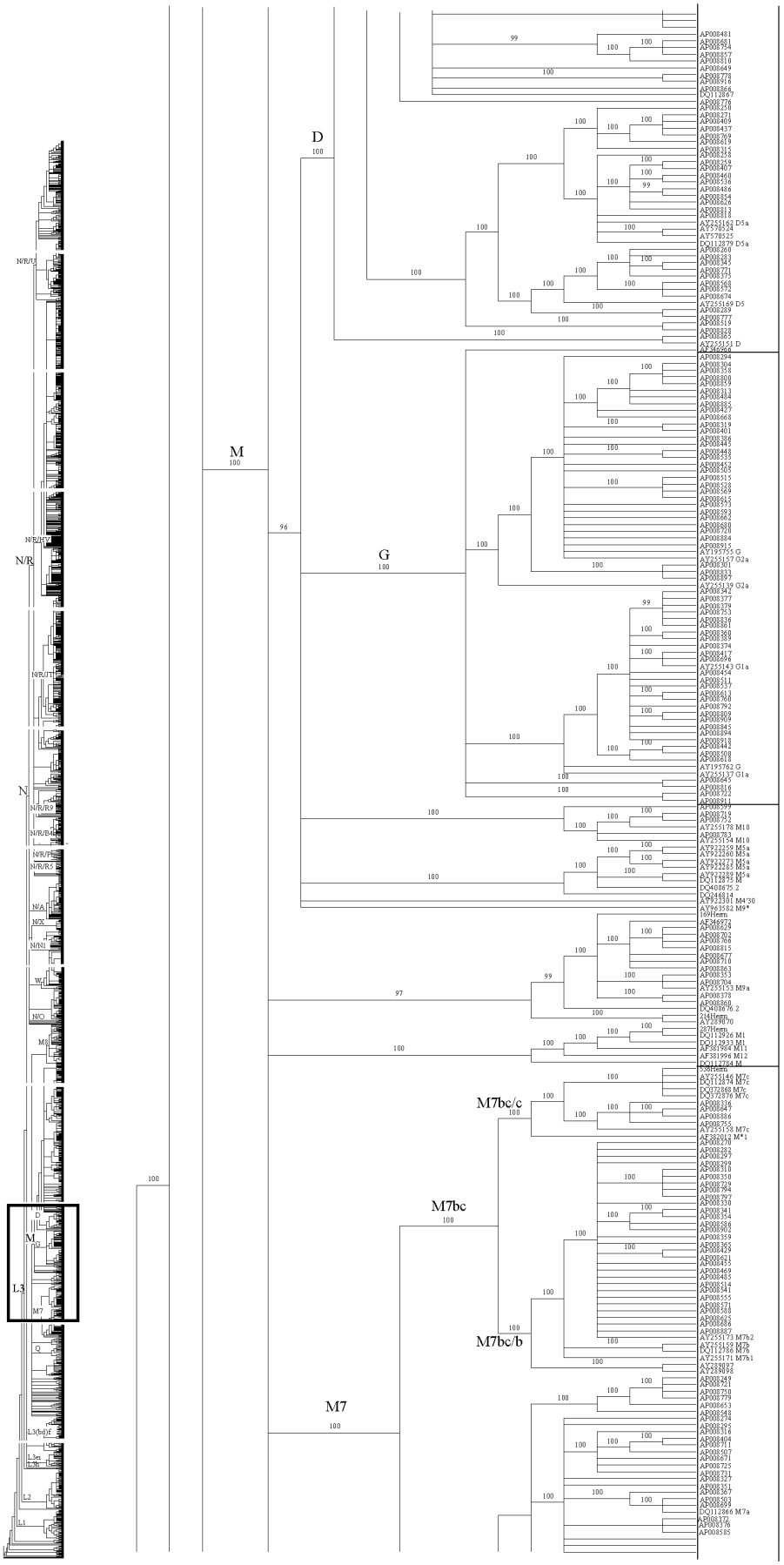

N/R/JT/T
N/R/JT/J
N/R
N/R/B5
N/R/JT/T
N/R/JT/J
N/R
N/R/B5
L3

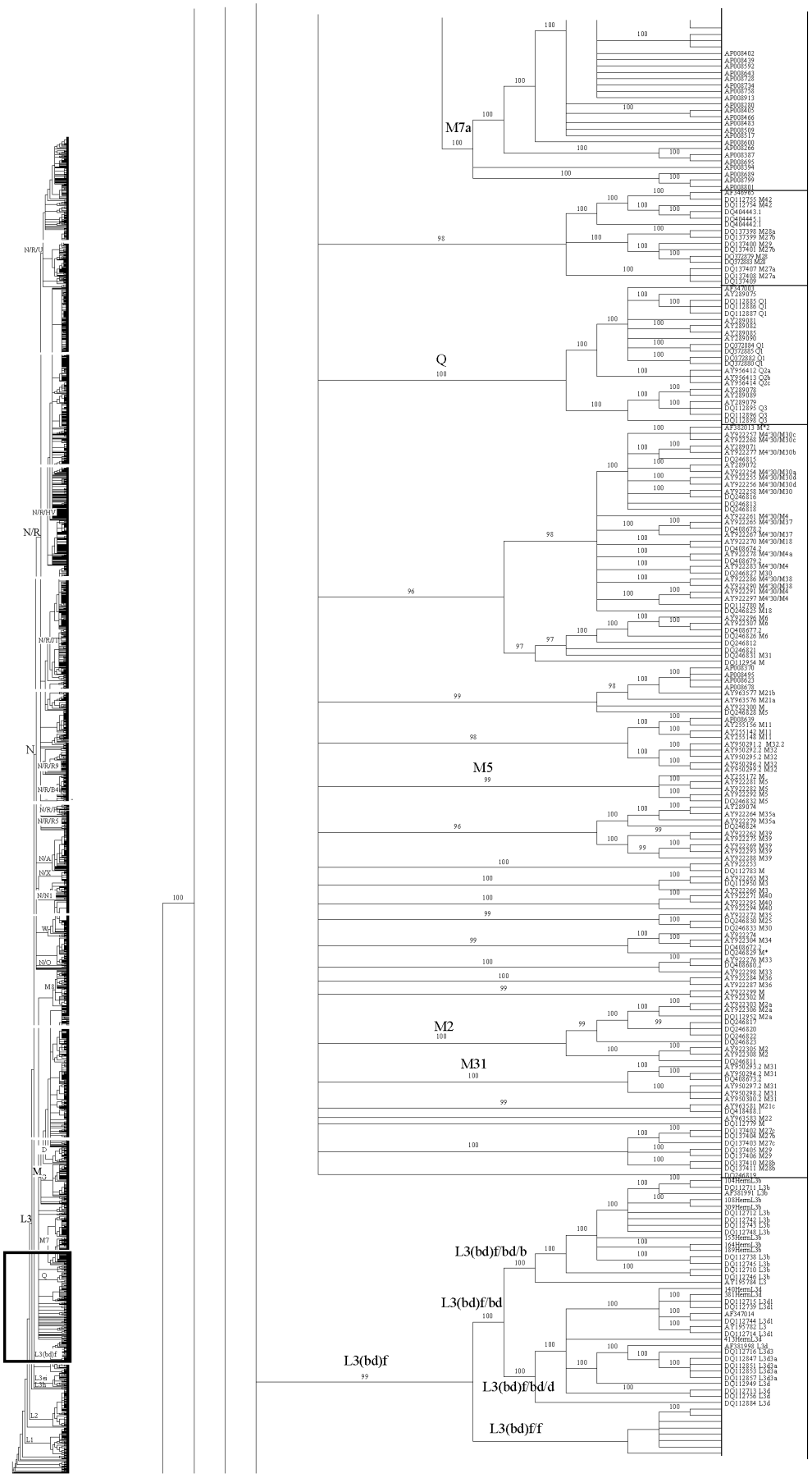

N/R/B5a
N/R/B5
N/R/B5b
N/R/B4bd
N/R/R9
N/R/B4
N/R/B5
N/R
N/R/B4bd
N/R
N/R/R9
N/R/B4
N/R/P

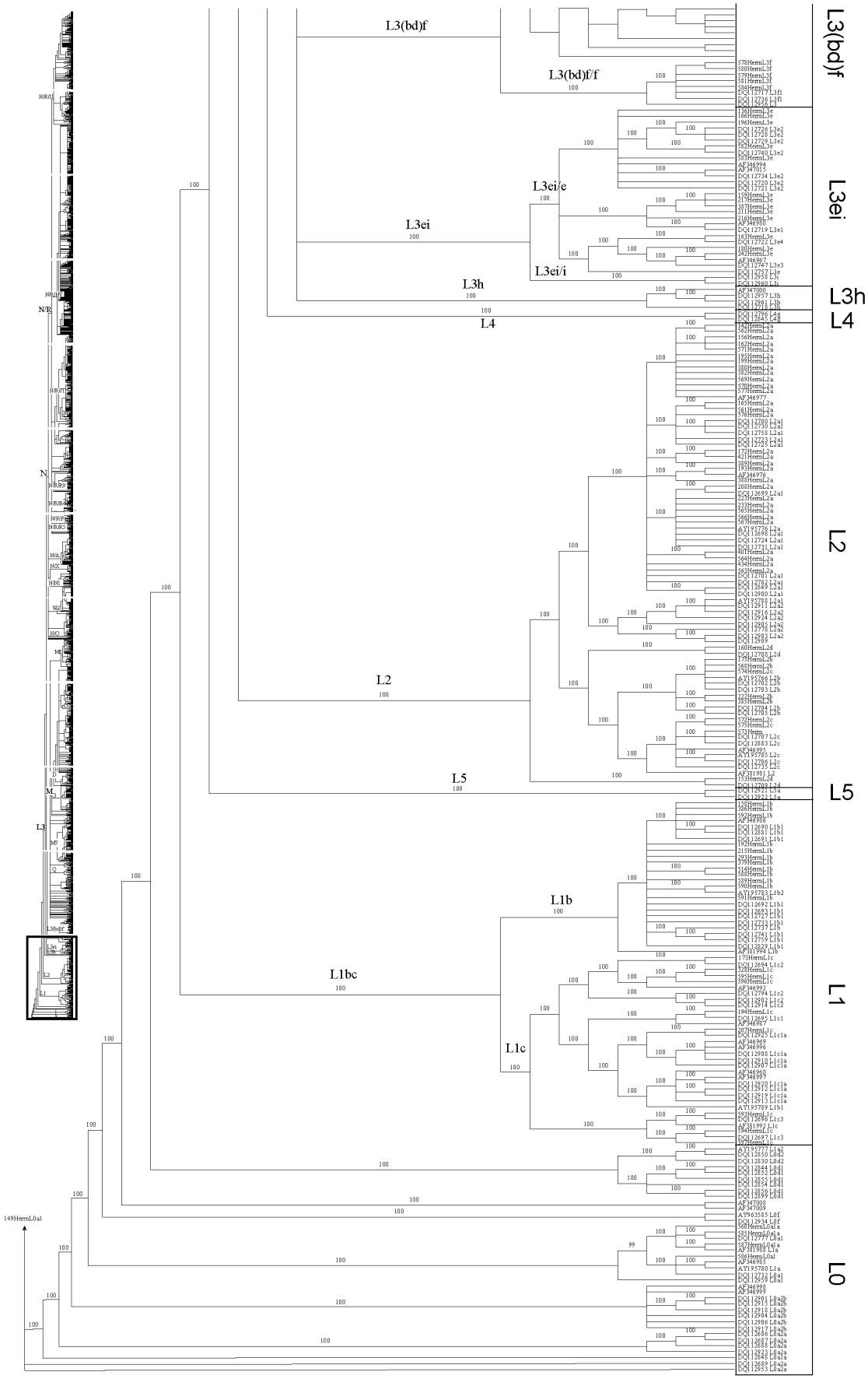

N/R/P
N/R
N/R/R5
N/A
N/X
N/N1
N/W
N/N1/I?


N/M
N
N/O
N
M/M8
M
M/D
M8/C
M8
M8/a
M8/Z
N/O
N/R/U
N/R/U1
N/R
N/R/J1
N
N/R/R9
N/R/B1
N/R/H2
N/R/K5
N/A1
N/X
N/N1
W1
N/O
M8
D
Q
M
L3
M7
Q
L3(b-d)f
L3ei
L3h
L2
L1


M/D


M/D
M/G
M
M/M7
D
G
M
M7
M7bc
M7bc/c
M7bc/b
L3
N/R/U
N/R
N/R/U7
N/R/U1
N
N/R/R9
N/R/B4
N/R/F
N/R/R5
N/A
N/X
N/N1
W1
N/O
M8
M
D
G
Q
L3
M7
Q
L3(b4)f
L3e1
L3b
L2
L1


M7a
Q
M2
M31
M5
L3(bd)f
L3(bd)f/bd/b
L3(bd)f/bd
L3(bd)f/bd/d
L3(bd)f/f
M/M7
M
M/Q
M
L3(bd)f


L3(bd)f
L3(bd)f/f
L3ei/e
L3ei
L3ei/i
L3h
L4
L2
L5
L1b
L1bc
L1c
149HermL0a1
L3(bd)f
L3ei
L3h
L4
L2
L5
L1
L0

Appendix E: Supplementary Tables

E1.1 Oceanic population size estimates 1930-2003 186
E1.2 Y-chromosome review table 187
E1.3 Oceanic mtDNA studies review table..... 190
E2.1 Polymorphism lists for mt genome sequences from this study 193
E5.1 Parsimony scores for characters from 75-taxon analysis 200
E6.1 Haplotype details HVR-I nt16065-nt16373 data set (Oceanic samples)..... 209
E6.2 Haplotype details HVR-I nt16189-nt16373 data set 215

E1.1 Oceanic population size estimates 1930-2003

Population size estimates for 18 Oceanic island groups from the World Data Analyst function in the Encyclopaedia Britannica® Online Academic Edition © 2007 Encyclopaedia Britannica, Inc.

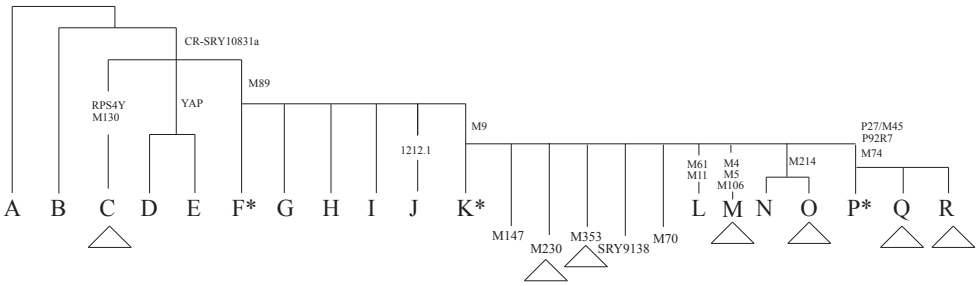
Year	Papua New Guinea	Solomon Islands	Vanuatu	New Caledonia	Northern Mariana Islands	Palau
1930	1306000	94000	36000	54000	19000	8000
1940	1308000	94000	43000	53000	48000	25000
1950	1412466	106647	52000	59000	6286	7251
1960	1746986	126363	65000	79000	8861	9482
1970	2288056	163000	86000	110000	12359	12005
1980	2991217	232452	116771	140386	16890	13311
1990	3758439	315139	146610	170900	43932	15207
2000	5186684	415939	189655	210965	69666	19237
2003	5582763	450000	204100	220358	73300	20240
Increase (%) 1930-2003	427.47	478.72	566.94	408.07	385.79	253.00

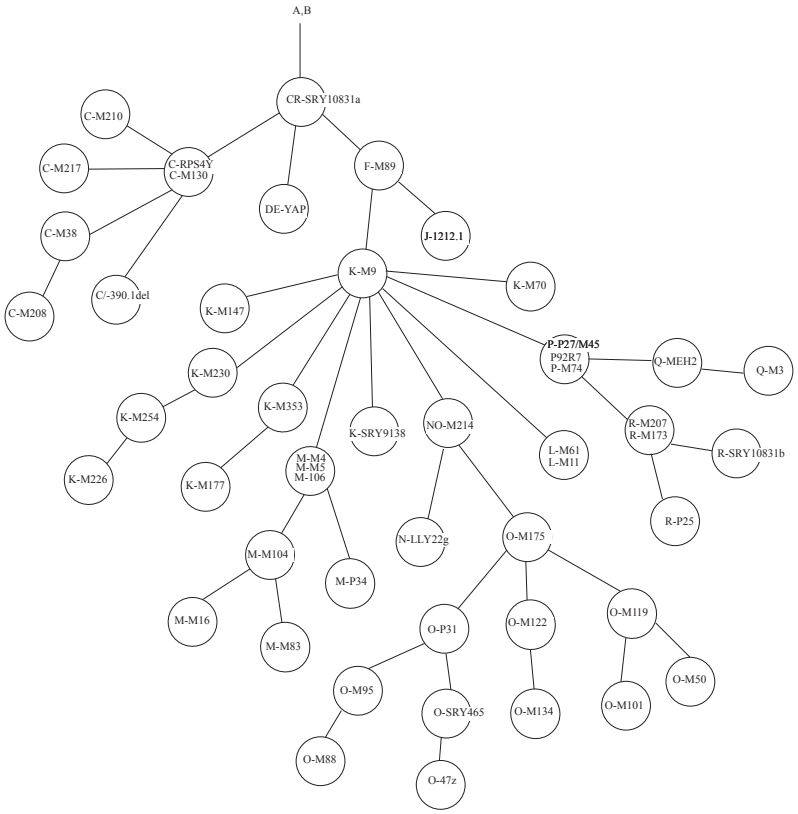
Year	Guam	Kiribati	Marshall Islands	Micronesia	Samoa	Tonga
1930	19000	27000	10000	32000	55000	28000
1940	22000	29000	10500	27000	61000	37000
1950	60000	33448	10904	32000	81858	50000
1960	67000	40732	15120	45000	110043	65000
1970	85000	48899	21714	61000	142331	80000
1980	106869	57536	30681	73000	155000	91694
1990	133653	71341	45830	96000	160000	95762
2000	155388	84183	53064	107411	174663	99900
2003	163000	87907	56429	111572	178800	101700
Increase (%) 1930-2003	857.89	325.58	564.29	348.66	325.09	363.21

Year	Tuvalu	American Samoa	Fiji	French Polynesia	Nauru	New Zealand
1930	4000	10000	181000	39000	3000	1491000
1940	4000	13000	218000	50000	3000	1637000
1950	4676	19100	289000	62000	3432	1909000
1960	5264	20000	394000	84000	4475	2377000
1970	5815	27267	520000	116700	6700	2820000
1980	7490	32418	634000	150900	7710	3144000
1990	9168	47011	737000	196700	9488	3452189
2000	10500	57582	810400	235191	11845	3858600
2003	10200	61194	826658	248272	12570	4001000
Increase (%) 1930-2003	255.00	611.94	456.72	636.59	419.00	268.34

E1.2 Y-chromosome review table

The outline structure of the human Y-chromosome tree is shown below, detailing major alphabetically named haplogroups (adapted from Jobling and Tyler-Smith 2003). Haplogroups are defined by one or more markers; generally two-state SNPs (single-nucleotide polymorphisms) which continue to be identified. These markers have alphanumeric names, many simply for the laboratory of origin; for example M3-M224 are from studies by Peter Underhill and colleagues (Y Chromosome Consortium: http://ycc.biosci.arizona.edu/nomenclature_system/frontpage.html, accessed 25/09/06). 'Paragroups' are labelled with an asterisk; such that sequences with an ancestral polymorphism at for example the marker M9 which defines groups K-R, but none of the subsequent variants found to date within this group are labeled K*; whereas an individual with the M9 variant and additional M4 marker is designated as belonging to haplogroup M. As new markers are found the tree is extended, and the resolution increases. More than 200 markers had been described in 2003, of which only 5 were believed to be recurrent (Jobling and Tyler-Smith 2003).
The labelled diagram of markers corresponds to the markers described in the compilation table of Pacific and neighbouring populations which follows on the next two pages. The markers used by each of the five studies included are shown in bold in the skeleton outlines below the table.





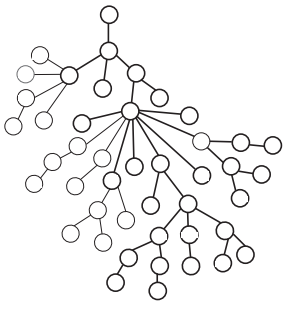
			A,B,DE			C								F*, F/J			F/K*					F/K/M										
Population	**Source**	**n**	Y*(xC-M130*, DE-Yap, F-M89*)	xC-RPS47, K-M9	DE-YAP	C-M130*	C-RPS4Y*(xC-M38*)	C-RPS4Y*(xC-M217, C/-390.1del, C-M38*, C-M210)	C-M217	C/-390.1del	C-M38*(xC-M208)	C-M38*	C-M208	F-M89*(xK-M9*)	F-M89*(xK-M9*, J-1212.1)	J-1212.1	K-M9*(xO-M175*, P-M45*, K-SRY9138, M-M4*, K-M178, K-M70, K-M11)	K-M9*(xO-M22*)	K-M9*(xL-M11, P-92R7*, NO-M214*, M-M106*, K-M177, K-M230*)	K-M9*(xK-M230*, K-M353*, M-M4*, NO-M214*, P-M74*)	K-M9*(xM-M4/5*, NO-M214*, P-P27/M45*, L-M61, K-M70, K-M147, K-SRY9138)	M-M106*	M-M5/M4*	M-M4*(xM-M104*, M-P34)	M-P34	M-M104*(xM-M16, M-M83)	NO-M214*(xO-M175*)	N-LLY22g	O-M175*	O-M175*(xO-M119*, O-M122*, O-M95*)	O-M175*(xO-P31*, O-M122*, OM119*)	O-M119*
Cook Islands	1,3,4	105	1	1		16					1		65	1				1		3										1		
Samoa	1,3,4	102				15	2				1		42	2				3		2	1				1	1						2
Tonga	3,4	37									1		12	1				2		1						5						
Maori	3,5	57										23	8				1	1														
East Futuna	4	50											15							7						5						
Tuvalu	4	100											17	2						36												1
Fiji	4	107						1			9		14	1						27				26		8	1					6
Vanuatu	1,2	286				50										1			95		30	69	7						10			
Niue	4	10																														
Tokelau	4	6											1							3												
Kapingamarangi	1	21				6															14											
Majuro	1	11													1	1					7											
Remote Oceania																																
count		***892***	***1***	***1***		***87***	***2***	***1***			***12***	***23***	***174***	***7***	***1***	***2***	***1***	***7***	***95***	***79***	***52***	***69***	***7***	***26***	***1***	***19***	***1***		***10***	***1***		***9***
% total R.O.			***0.11***	***0.11***		***9.75***	***0.22***	***0.11***			***1.35***	***2.58***	***19.51***	***0.78***	***0.11***	***0.22***	***0.11***	***0.78***	***10.65***	***8.86***	***5.83***	***7.74***	***0.78***	***2.91***	***0.11***	***2.13***	***0.11***		***1.12***	***0.11***		***1.01***
Papua New Guinea	1,4	189				2		1			5		5		1					7	9		28	1	64	2						1
West New Guinea	4	183									8		23							10				6	133							
Trobriand Islands	4	53											5							12				1	15							15
Solomon Islands	2	32																	19			3							9			
New Britain	4	19									3		1							6					1	5						
Near Oceania																																
count		***476***				***2***		***1***			***16***		***34***		***1***				***19***	***35***	***9***	***3***	***28***	***8***	***213***	***7***			***9***			***16***
% total N.O.						***0.42***		***0.21***			***3.36***		***7.14***		***0.21***				***3.99***	***7.35***	***1.89***	***0.63***	***5.88***	***1.68***	***44.75***	***1.47***			***1.89***			***3.36***
Philippines	1,4	67						4						1						1	1									1		16
Borneo	1,4	127				15			1		1			2	1	2				2	1						2	1		2		6
Moluccas	4	34						1			7									6					7							2
Nusa Tenggaras	4	31						2			5									10					2							7
Java	4	53						1						1						1										1		12
Sumatra	4	57			1			3						8						2												10
Malaysia	4	18							2					1						2												2
Island SEA																																
count		***387***			***1***	***15***		***11***	***3***		***13***			***13***	***1***	***2***				***24***	***2***				***9***		***2***	***1***		***4***		***55***
% total Island SEA					***0.26***	***3.88***		***2.84***	***0.78***		***3.36***			***3.36***	***0.26***	***0.52***				***6.20***	***0.52***				***2.33***		***0.52***	***0.26***		***1.03***		***14.21***
Taiwan	1,4	82																												2	1	34
China	4	36							2																		4					8
Taiwan (Chinese)	4	26							1																		1			3		6
Korea	4	25							3					2													2			8		1
Vietnam	4	10							1																							1
Taiwan & Asia																																
count		***179***							***7***					***2***													***7***			***13***	***1***	***50***
total Taiwan & Asia									***3.91***					***1.12***													***3.91***			***7.26***	***0.56***	***27.93***
Sandy Desert	4	35								24				1						6												
Arnhem Land	4	60						6		32				1						18												
Australia																																
count		***95***						***6***		***56***				***2***						***24***												
% total Australia								***6.32***		***58.95***				***2.11***						***25.26***												

Table E1.2 Y chromosome review table

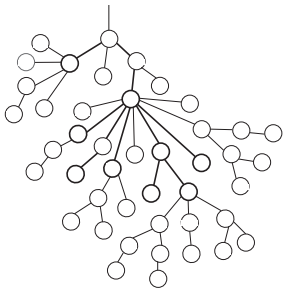
Study sources: 1. Hurles et al 2005, 2. Cox and Lahr 2006, 3. Fris 2006, 4. Kayser et al 2006, 5. Underhill et al 2001. Data has been combined for some localities; for example several subsets - coast, highlands, Bereina, Kapuna, are combined as 'Papua New Guinea' and 'Borneo' contains both the Benjarmasin and the Kota Kinabalu sample from Hurles et al 2005. An asterisk following a marker denotes the haplotype belongs within a 'paragroup' from which other known sublineages descend, and where a marker is shown prefaced with an 'x', this means it has been tested and the ancestral form found. For example F-M89*(xK-M98) means the haplotype has the derived M89 SNP allele, but not the K-defining M98 derived allele, and therefore falls within the F paragroup.

Right: The markers examined in each of the five studies reviewed vary in coverage and these are summarised in the skeleton outlines of the detailed SNP marker tree (previous page) shown at right.

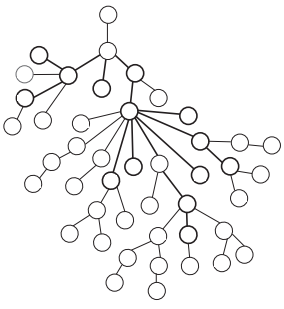
F/K/N-O																		F/K/P				F/K/M230					F/K/M353	
NO-M214*(xO-M175*)	N-LLY22g	O-M175*	O-M175*(xO-M119*, O-M122*, O-M95*)	O-M175*(xO-P31*, O-M122*, OM119*)	O-M119*	O-M119*(xO-M50, O-M101)	O-M101	O-M50	O-M122*	O-M122*(xO-M134)	O-M134	O-P31*(xO-M95*, O-SRY465*)	O-M95*	O-M95*(xO-M88)	O-M88	P-M45*(xR-M173*, P-M120, P-M125, P-M24)	P-M74*(xR-M173*)	R-M207*	R-M207*(xR-SRY10831b, R-P25)	R-M173*	R-M173*(xR-M17)	K-M230*	K-M230*(xK-M254*)	K-M254*(xK-M226)	K-M226	K-M353*(xK-M177)	**Population**
			1		2	1			5	7 18	2		1	1						8 1				1			Cook Islands Samoa
									3 5	12						1					18						Tonga Maori
					1					21 41			3							1 2						1 4	East Futuna Tuvalu Fiji
1		10			6			1		8 4								4				15					Vanuatu
										9 2										1							Niue Tokelau
														1													Kapingamarangi
								2																			Majuro
																											Remote Oceania
1 ***0.11***		***10*** ***1.12***	***1*** ***0.11***		***9*** ***1.01***	***1*** ***0.11***		***3*** ***0.34***	***13*** ***1.46***	***122*** ***13.68***	***2*** ***0.22***		***4*** ***0.45***	***2*** ***0.22***		***1*** ***0.11***		***4*** ***0.45***		***13*** ***1.46***	***18*** ***2.02***	***15*** ***1.68***		***1*** ***0.11***		***5*** ***0.56***	***count***
																											% total R.O.
					1					16							1						1	43	2		Papua New Guinea
					15					1 5														2			West New Guinea Trobriand Islands
		9																				1					Solomon Islands
										1														2			New Britain
																											Near Oceania
		9 ***1.89***			***16*** ***3.36***					***23*** ***4.83***							***1*** ***0.21***					***1*** ***0.21***	***1*** ***0.21***	***47*** ***9.87***	***2*** ***0.42***		***count***
																											% total N.O.
			1		16	9		1		28	3		1							1							Philippines
2	1		2		6	15	1	8		18	4		15	21	3		1		4	1							Borneo
					2					4														7			Moluccas
					7					1										1			1	2			Nusa Tenggaras
			1		12					11	1		22				1			2							Java
					10					16	1		8				2			4				2			Sumatra
					2					3	2		5							1							Malaysia
																											Island SEA
2 ***0.52***	***1*** ***0.26***		***4*** ***1.03***		***55*** ***14.21***	***24*** ***6.20***	***1*** ***0.26***	***9*** ***2.33***		***81*** ***20.93***	***11*** ***2.84***		***51*** ***13.18***	***21*** ***5.43***	***3*** ***0.78***		***4*** ***1.03***		***4*** ***1.03***	***10*** ***2.58***				***1*** ***0.26***	***11*** ***2.84***			***count***
																											% total Island SEA
			2	1	34	15		13		9	3	1	2		2												Taiwan
4					8					14	7		1														China
1			3		6					7	8																Taiwan (Chinese)
2			8		1					6	1						1			1							Korea
					1					4			4														Vietnam
																											Taiwan & Asia
7 ***3.91***			***13*** ***7.26***	***1*** ***0.56***	***50*** ***27.93***	***15*** ***8.38***		***13*** ***7.26***		***40*** ***22.35***	***19*** ***10.61***	***1*** ***0.56***	***7*** ***3.91***		***2*** ***1.12***		***1*** ***0.56***			***1*** ***0.56***							***count***
																											% total Taiwan & Asia
											1									3							Sandy Desert
																				3							Arnhem Land
																											Australia
											1 ***1.05***									***6*** ***6.32***							***count***
																											% total Australia



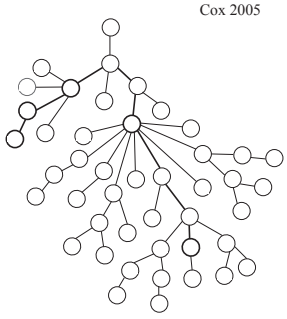
Hurles et al 2005



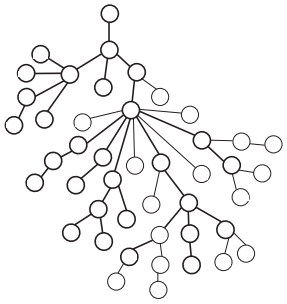
Cox 2005



Underhill et al 2001



Fris 2006



Kayser et al 2006

E1.3 Oceanic mtDNA studies review table

Year	Authors	Samples	Target	Notes	Entrez database
1989	M. Hertzberg, K.N. Mickleson, S.W. Serjeantson, et al., *Am J Hum*	Oceanic and Australian populations (n= 307)	Testing for 9-bp deletion	Conclude pre-Polynesians 'were ultimately derived from east Asia'	n/a
1992	S.W. Ballinger, T.G. Schurr, A. Torroni, et al., *Genetics*	153 samples from East and Southeast Asia and Island SEA	RFLPs	'Southern Mongoloid origin of Asians', 'remnants of the founding population of Papua New Guinea were found in Malaysia'	n/a
1993	E. Hagelberg and J.B. Clegg, *Proc Biol Sci*	21 bone samples from Oceania, dating from ~2500BP-~200BP	9bp deletion typing and some HVR-I sequencing	Early Lapita samples from Watom and Fiji did not have the 9bp deletion. Later remains from Tonga, NZ, the Chathams, Society Islands and Hawaii did.	N
1994	E. Hagelberg, S. Quevedo, D. Turbon, et al., *Nature*	12 bone samples from Easter island; dating to between ~850BP and 100BP	9bp deletion typing and HVR-I	All samples have 9bp deletion and PM, with one having additional 16271T and two sharing additional 16292T.	N
1994	J.K. Lum, O. Rickards, C. Ching, et al., *Hum Biol*	197 individuals, Oceania, Australia, Island and mainland Asia (plus African)	9bp del typing, 76 HVR-I sequences	Defines group I [B], group II [Q] and group III [F?16294, 16304, 16362]	?incl with later (2000 accessions
1995	T. Melton, R. Peterson, A.J. Redd, et al., *Am J Hum Genet*	1037 samples from 12 populations from Asia, Taiwan and Island Southeast Asia	9bp and SSO typing	Find results consistent with ancestry of the B4a type in Taiwan with migration to Island Southeast Asia and the Pacific.	N
1995	A.J. Redd, N. Takezaki, S.T. Sherry, et al., *Mol Biol Evol*	74 samples from Indonesia, coastal PNG and American Samoa, all with 9bp del. Also sequenced additional 35 without deletion	Entire control region	Conclude 9bp deletion arose independently in Asia and in Africa, and possibly multiple times in Asia; increase in frequency of the B4a type moving eastwards associated with a decrease in diversity, consistent with founder events.	Y
1995	B. Sykes, A. Leiboff, J. Low-Beer, et al., *Am J Hum Genet*	Oceania, Island SEA and Taiwan, n=1178	9bp del screening, screening for 16265, some HVR-I seq'ing	Defined three lineage groups: 9bp del [B], 16265 transversion [Q], and one with 16172 and 16304 [F?]variants.	Y (haps)
1996	D.J. Betty, A.N. Chin-Atkins, L. Croft, et al., *Am J Hum Genet*	310 Australian samples	9bp deletion screening	Multiple origins of the 9bp deletion in Australia	n/a
1998	T. Melton, S. Clifford, J. Martinson, et al., *Am J Hum Genet*	Taiwan (n=28)	HVR-I and II sequences	Conclusion that Taiwanese have common ancestry with mainland Asian populations, but have undergone period of isolation. Notable to me - their sample AMI2 has same sequence as AMI15 (M7c with 9bp deletion).	N
1998	M. Richards, S. Oppenheimer and B. Sykes, *Am J Hum Genet*	Analysis of existing data Oceania, Taiwan and ISEA	Calculate divergence times using Forster et al 1996 method and rate.	Conclude ancestry of the Polynesian motif is dated to the Pleistocene (~17,000 years ago), and likely location of ancestor is Southeast Asia	n/a
1998	R.P. Murray-McIntosh, B.J. Scrimshaw, P.J. Hatfield, et al., *Proc Natl Acad Sci U S A*	31 NZ Maori	HVR-I, and founding pop. size analysis	Estimate ~50-100 founding women	N
1998	S. van Holst Pellekaan, M. Frommer, J. Sved, et al., *Am J Hum Genet*	114 Australian samples	HVR-I		Y (haps)

Year	Authors	Samples	Target	Notes	Entrez database
1998	J.K. Lum and R.L. Cann, *Am J Phys Anthropol*	Oceania, Australia, Island Southeast Asia, n=973	9 bp del. polymorphisms, distance analyses on this and language and geography	Suggest extensive gene flow throughout Micronesia but substantial isolation in other Pacific regions.	n/a
1998	J.K. Lum, R.L. Cann, J.J. Martinson, et al., *Am J Hum Genet*	Samples as for Lum et al 1998 AJPA	HVR-I and nuclear STRs	Results consistent with initial settlement of Remote Oceania from island Southeast Asia , and with extensive postcolonisation male-biased gene flow with Near Oceania	?with later (2000) accessions
1999	D.A. Merriwether, J.S. Friedlaender, J. Mediavilla, et al., *Am J Phys Anthropol*	Large sample set – typed 1800 samples	9bp deletion and RFLP for nts 10398, 10400, 16398 and 16519.	Earlier interpretation of B4a lineages in Island Melanesia: views deletion as 'Austronesian' indicator. Later papers from this group find more complex pattern between language and mtDNA distributions.	N
1999	A.J. Redd and M. Stoneking, *American Journal of Human Genetics*	Australia, east Indonesia and PNG.	9bp typing, SSO typing. Control region sequencing for Australian (n=53) and 25 east Indonesian samples.	Results do not support close ancestral relationships between PNG and Australian populations	Y
2000	M. Ingman, H. Kaessmann, S. Paabo, et al., *Nature*	Included 1 Samoan and 5 PNG samples in a global dataset	Whole mtDNA sequencing		Y
2000	J.K. Lum and R.L. Cann, *Am J Phys Anthropol*	Samples as for Lum et al 1994, Lum et al 1998, new sequence data for 116 samples, and from databases	9 bp del screen, HVR-I (from 16189)	Marianas and Yap settled directly from Island SEA, receiving migrants from rest of Micronesia subsequently, genetic similarities among Micronesians and Polynesians likely result of combination of common origin and gene flow	Y
2001	P. Forster, A. Torroni, C. Renfrew, et al., *Mol Biol Evol*	Existing datasets	RFLPs, star contraction method	Global conclusions eg out-of Africa, post-ice age expansions	n/a
2002	J.S. Friedlaender, F. Gentz, K. Green, et al.,	Santa Cruz Islands (n=64)	HVR-I and HVR-II	Found B4a, Q, P and M28 haplotypes	N
2002	M. Tommaseo-Ponzetta, M. Attimonelli, M. De Robertis, et al., *Am J Phys* Anthropol	202 individuals from Irian Jaya	HVR-I and 9bp del screen	No 9bp del samples found; conclude prolonged isolation	N
2003	M. Ingman and U. Gyllensten, *Genome Res*	51 mt genome sequences analysed with 2000 dataset: includes 20 Australian, 21 PNG, and 6 from	Whole mtDNA sequencing	High diversity in the Australian sequences, and some shared clades between PNG and Australia suggesting colonisation from a common source and/or later mixing	Y
2003	M.P. Cox, PhD thesis, University of Otago	~600 individuals from Vanuatu, Indonesia and Madagascar	mtDNA (control plus coding SNPs) and Y-chromosome markers (RFLPs used)	Earlier non-Austronesian peoples contributed substantially to present day populations in Indonesia and Vanuatu	N
2005	J.S. Friedlaender, F. Gentz, F.R. Friedlaender, et al., *Papuan Pasts: Linguistic, Archaeological and Biological Relations of Papuan Peoples*	Large sample (n=886), Near Guinea, Island Melanesia, with n=54 Remote Oceania.	HVR-I and II, RFLP typing nts 10398 and 10400.	Complex relationship between mtDNA and languages in Island Melanesia	N

Year	Authors	Samples	Target	Notes	Entrez database
2005	J. Friedlaender, T. Schurr, F. Gentz, et al., *Mol Biol Evol*	Bismarck Archipelago and Bougainville (n=886)	HVR-I and HVR-II, subsequent RFLPs coding and whole genome seq'ing 3 Q2 individuals	P and Q networks suggest ancient population expansions following first settlement, subsequent expansion of Q2 in Island Melanesia approx 10,000 years later. Localised distributions of haplotypes (both P and Q) suggest highly restricted female movement.	3 mt genome seqs on database, not control region sequences
2005	D.A. Merriwether, J.A. Hodgson, F.R. Friedlaender, et al., *Proc Natl Acad Sci U S A*	Same dataset as Friedlander et al 2005	14 whole mt sequences	Defines haplogroups M27, M28 and M29	Y
2005	A.L.H. Whyte, S.J. Marshall, G.K. Chambers, *Human Biology*	NZ Polynesian: 61 Maori, 24 Eastern Polynesian	Portion HVR-I: 16189-16360.	Propose larger founding population for New Zealand ~190 (170-230) than Murray-McIntosh et al 1998.	Y
2005	M.P. Cox, *Human Biology*	Existing datasets (Murray's PhD)	Repeats dating of PM as in Richards et al 1998	Concludes that evidence consistent with southern China/Taiwan origin of proto-PM, not the 17,000 East Indonesian result of Richards et al (1998).	Y (2 haps)
2005	J.A. Trejaut, T. Kivisild, J.H. Loo, et al., *PLoS Biology*	Taiwan: 640 individuals	HVR-I, markers and whole mts for 8 samples	Defines subclade B4a1a	Y
2006	J. Ohashi, I. Naka, K. Tokunaga, et al., *Am J Phys Anthropol*	Balopa Islands, Manus Province PNG (n=59), Gidra-speakers (NAN) from southwestern PNG (n=59), Tonga	9bp del typing and HVR-I sequencing		Y
2006	S.M. van Holst Pellekaan, M. Ingman, J. Roberts-Thomson, et al., *Am J Phys Anthropol*	49 samples from Australia	8 whole mtDNAs, 41 partial sequences	Deep lineages present in Australia branching from both M and N	Y (whole mts only)

E2.1 Polymorphism lists for mt genome sequences from this study

Polymorphisms which are 'private': found only in a single individual in the base-labelled phylogenies shown in Chapter 3 are highlighted in bold. All chromatograms were edited manually using Sequencher™ (Gene Codes Corporation), and private polymorphisms were carefully re-checked. Four of the twenty samples had polymorphisms which were unusual in a phylogenetic sense (requiring back-mutations in a phylogeny, or homoplasious). Further details of these changes are follow the polymorphism lists for the mt genomes.

DQ372868 AMI15

52 differences to rCRS. Haplogroup M/M7

Base	AMI15	rCRS
73	G	A
146	C	T
152	**C**	**T**
199	C	T
263	G	A
309.1	C	:
315.1	C	:
489	C	T
514	:	C
515	:	A
750	G	A
1,438	G	A
2,706	G	A
3,107	:	C
3,606	G	A
4,071	T	C
4,769	G	A
4,850	T	C
5,442	C	T
6,455	T	C
7,028	T	C
8,271	**:**	**A**
8,272	**:**	**C**
8,273	**:**	**C**
8,274	**:**	**C**
8,275	**:**	**C**
8,276	**:**	**C**
8,277	**:**	**T**
8,278	**:**	**C**
8,279	**:**	**T**
8,701	G	A
8,860	G	A
9,540	C	T
9,824	C	T
10,398	G	A
10,400	T	C
10,873	C	T
11,665	T	C
11,719	A	G
12,091	C	T
12,705	T	C
14,766	T	C
14,783	C	T
15,043	A	G
15,236	G	A
15,301	A	G
15,326	G	A
16,129	**A**	**G**
16,223	T	C
16,295	T	C
16,362	C	T
16,519	C	T

DQ372869 PAI9

47 differences to rCRS. Haplogroup N/R/B5a

Base	PAI9	Seq2
73	G	A
93	G	A
204	**C**	**T**
210	G	A
263	G	A
315.1	C	:
514	:	C
515	:	A
709	A	G
750	G	A
1,438	G	A
2,706	G	A
3,107	:	C
3,537	G	A
4,769	G	A
7,028	T	C
8,271	:	A
8,272	:	C
8,273	:	C
8,274	:	C
8,275	:	C
8,276	:	C
8,277	:	T
8,278	:	C
8,279	:	T
8,557	**A**	**G**
8,584	A	G
8,614	**C**	**T**
8,860	G	A
9,950	C	T
9,962	A	G
10,398	G	A
11,149	A	G
11,151	T	C
11,719	A	G
14,149	T	C
14,766	T	C
15,046	**G**	**A**
15,235	G	A
15,301	**A**	**G**
15,326	G	A
16,140	C	T
16,183	C	A
16,189	C	T
16,266	**G**	**C**
16,362	**C**	**T**
16,519	C	T

DQ372870 TRO122

24 differences to rCRS. Haplogroup N/R/P2

Base	TRO122	rCRS
73	G	A
152	C	T
263	G	A
309.1	C	:
315.1	C	:
750	G	A
2,706	G	A
3,107	:	C
3,203	G	A
3,882	A	G
4,122	G	A
4,769	G	A
5,423	G	A
7,028	T	C
8,572	A	G
8,859	T	C
8,860	G	A
11,719	A	G
11,914	A	G
14,766	T	C
14,890	G	A
15,326	G	A
15,607	G	A
16,519	C	T

DQ372871 TRO131

38 differences to rCRS. Haplogroup N/R/B4a1a

Base	TRO131	rCRS
73	G	A
146	C	T
263	G	A
309.1	C	:
315.1	C	:
514	:	C
515	:	A
750	G	A
1,438	G	A
2,706	G	A
3,107	:	C
4,769	G	A
5,465	C	T
6,719	C	T
7,028	T	C
8,271	:	A
8,272	:	C
8,273	:	C
8,274	:	C
8,275	:	C
8,276	:	C
8,277	:	T
8,278	:	C
8,279	:	T
8,860	G	A
9,123	A	G
10,238	C	T
11,719	A	G
12,239	T	C
14,766	T	C
15,326	G	A
15,746	G	A
16,182	C	A
16,183	C	A
16,189	C	T
16,217	C	T
16,261	T	C
16,519	C	T

DQ372872 TRO133

25 differences to rCRS. Haplogroup N/R/P2

Base	TRO133	rCRS
73	G	A
152	C	T
263	G	A
309.1	C	:
315.1	C	:
750	G	A
2,706	G	A
3,107	:	C
3,203	G	A
3,882	A	G
4,122	G	A
4,769	G	A
5,423	G	A
7,028	T	C

8,572	A	G
8,859	T	C
8,860	G	A
11,719	A	G
11,914	A	G
12,026	**G**	**A**
14,766	T	C
14,890	G	A
15,326	G	A
15,607	G	A
16,519	C	T

DQ372873 TRO137
41 differences to rCRS. Haplogroup N/R/B4a1a1

Base	TRO137	rCRS
73	G	A
146	C	T
263	G	A
315.1	C	:
449	**C**	**T**
514	:	C
515	:	A
750	G	A
1,438	G	A
2,706	G	A
3,107	:	C
4,769	G	A
5,465	C	T
6,253	**C**	**T**
6,719	C	T
7,028	T	C
8,271	:	A
8,272	:	C
8,273	:	C
8,274	:	C
8,275	:	C
8,276	:	C
8,277	:	T
8,278	:	C
8,279	:	T
8,860	G	A
9,123	A	G
10,238	C	T
11,719	A	G
12,239	T	C
14,022	G	A
14,766	T	C
15,326	G	A
15,746	G	A
16,129	**A**	**G**
16,182	C	A
16,183	C	A
16,189	C	T
16,217	C	T
16,261	T	C
16,519	C	T

DQ372874 KAP19
39 differences to rCRS. Haplogroup N/R/B4a1a1

Base	KAP19	Seq2
73	G	A
146	C	T
263	G	A
302	C	A
514	:	C
515	:	A
750	G	A
1,438	G	A
2,706	G	A
3,107	:	C
4,769	G	A
5,465	C	T
6,719	C	T
7,028	T	C
8,271	:	A
8,272	:	C
8,273	:	C
8,274	:	C
8,275	:	C
8,276	:	C
8,277	:	T
8,278	:	C
8,279	:	T
8,860	G	A
9,123	A	G
10,238	C	T
11,719	A	G
12,239	T	C
14,022	G	A
14,766	T	C
15,326	G	A
15,746	G	A
15,924	G	A
16,182	C	A
16,183	C	A
16,189	C	T
16,217	C	T
16,261	T	C
16,519	C	T

DQ372875 KAP89
42 differences to rCRS. Haplogroup N/R/B4a1a1

Base	KAP89	rCRS
73	G	A
146	C	T
263	G	A
302	C	A
309.1	C	:
309.2	C	:
315.1	C	:
514	:	C
515	:	A
750	G	A
1,438	G	A
2,706	G	A
3,107	:	C
4,769	G	A
5,465	C	T
6,719	C	T
7,028	T	C
8,271	:	A
8,272	:	C
8,273	:	C
8,274	:	C
8,275	:	C
8,276	:	C
8,277	:	T
8,278	:	C
8,279	:	T
8,860	G	A
9,123	A	G
10,238	C	T
11,719	A	G
12,239	T	C
14,022	G	A
14,766	T	C
15,326	G	A
15,746	G	A
15,924	G	A
16,182	C	A
16,183	C	A
16,189	C	T
16,217	C	T
16,261	T	C
16,519	C	T

DQ372876 MJ22
43 differences to rCRS. Haplogroup M/M7c

Base	MJ22	rCRS
73	G	A
146	C	T
199	C	T
263	G	A
309.1	C	:
309.2	C	:
315.1	C	:
489	C	T
514	:	C
515	:	A
750	G	A
1,438	G	A
2,706	G	A
3,107	:	C
3,606	G	A
4,071	T	C
4,769	G	A
4,850	T	C
5,442	C	T
6,455	T	C
7,028	T	C
8,701	G	A
8,860	G	A
9,540	C	T
9,824	C	T
10,398	G	A
10,400	T	C
10,873	C	T
11,665	T	C
11,719	A	G
12,091	C	T
12,579	**G**	**A**
12,705	T	C
14,766	T	C
14,783	C	T
15,043	A	G
15,236	G	A
15,301	A	G
15,326	G	A
16,223	T	C
16,295	T	C
16,362	C	T
16,519	C	T

DQ372877 MJ86
42 differences to rCRS. Haplogroup N/R/B4a1a1

Base	MJ86	rCRS

73	G	A
146	C	T
195	**C**	**T**
263	G	A
309.1	C	:
315.1	C	:
514	:	C
515	:	A
750	G	A
1,438	G	A
2,706	G	A
3,107	:	C
4,769	G	A
5,465	C	T
6,216	**C**	**T**
6,719	C	T
7,028	T	C
8,271	:	A
8,272	:	C
8,273	:	C
8,274	:	C
8,275	:	C
8,276	:	C
8,277	:	T
8,278	:	C
8,279	:	T
8,860	G	A
9,123	A	G
10,238	C	T
11,719	A	G
12,239	T	C
14,022	G	A
14,766	T	C
15,326	G	A
15,746	G	A
15,924	G	A
16,182	C	A
16,183	C	A
16,189	C	T
16,217	C	T
16,261	T	C
16,519	C	T

DQ372878 PO314
41 differences to rCRS. Haplogroup N/R/B4a1a1/Pol. motif

Base	PO314	rCRS
73	G	A
146	C	T
263	G	A
310	**C**	**T**
514	:	C
515	:	A
750	G	A
1,438	G	A
2,706	G	A
3,107	:	C
4,769	G	A
5,465	C	T
6,719	C	T
7,028	T	C
8,271	:	A
8,272	:	C
8,273	:	C
8,274	:	C
8,275	:	C
8,276	:	C
8,277	:	T
8,278	:	C
8,279	:	T
8,860	G	A
9,123	A	G
10,238	C	T
10,484	**T**	**C**
11,719	A	G
12,239	T	C
14,022	G	A
14,766	T	C
15,326	G	A
15,746	G	A
16,182	C	A
16,183	C	A
16,189	C	T
16,217	C	T
16,247	G	A
16,261	T	C
16,519	C	T

DQ372879 PO332
45 differences to rCRS. HaplogroupM/M28

Base	PO332	rCRS
73	G	A
152	C	T
195	C	T
263	G	A
315.1	C	:
489	C	T
513.1	**C**	**:**
513.2	**A**	**:**
513.3	**C**	**:**
513.4	**A**	**:**
750	G	A
1,438	G	A
1,598	A	G
1,719	A	G
2,706	G	A
3,107	:	C
4,769	G	A
6,281	G	A
6,374	C	T
6,962	A	G
7,028	T	C
7,394	**G**	**A**
8,701	G	A
8,860	G	A
9,540	C	T
10,245	C	T
10,398	G	A
10,400	T	C
10,658	G	A
10,873	C	T
11,719	A	G
12,705	T	C
14,766	T	C
14,783	C	T
15,043	A	G
15,067	C	T
15,301	A	G
15,326	G	A
16,086	C	T
16,129	A	G
16,148	T	C
16,223	T	C
16,362	C	T
16,366	**T**	**C**
16,468	C	T

DQ372880 PO392
47 differences to rCRS. Haplogroup M/Q1

Base	PO392	rCRS
59	**C**	**T**
73	G	A
89	C	T
146	C	T
207	A	G
263	G	A
315.1	C	:
489	C	T
750	G	A
1,375	T	C
1,438	G	A
2,706	G	A
3,107	:	C
4,117	C	T
4,769	G	A
5,460	A	G
5,843	G	A
7,028	T	C
7,993	C	T
8,701	G	A
8,790	A	G
8,860	G	A
8,964	T	C
9,540	C	T
10,256	C	T
10,398	G	A
10,400	T	C
10,873	C	T
11,314	G	A
11,719	A	G
12,705	T	C
12,940	A	G
13,500	C	T
14,025	C	T
14,766	T	C
14,783	C	T
15,043	A	G
15,301	A	G
15,326	G	A
16,129	A	G
16,144	C	T
16,148	T	C
16,172	C	T
16,265	C	A
16,311	C	T
16,343	G	A
16,519	C	T

DQ372881 MF025
41 differences to rCRS. Haplogroup N/R/B4a1a1/Pol. motif

Base	MF025	rCRS
73	G	A
146	C	T
215	**G**	**A**
263	G	A
315.1	C	:
514	:	C
515	:	A
750	G	A
1,438	G	A

2,706	G	A
3,107	:	C
4,769	G	A
5,465	C	T
6,719	C	T
7,028	T	C
8,271	:	A
8,272	:	C
8,273	:	C
8,274	:	C
8,275	:	C
8,276	:	C
8,277	:	T
8,278	:	C
8,279	:	T
8,860	G	A
9,123	A	G
10,238	C	T
10,529	**G**	**A**
11,719	A	G
12,239	T	C
14,022	G	A
14,766	T	C
15,326	G	A
15,746	G	A
16,182	C	A
16,183	C	A
16,189	C	T
16,217	C	T
16,247	G	A
16,261	T	C
16,519	C	T

DQ372882 MO304

47 differences to rCRS. Haplogroup M/Q1

Base	MO304	rCRS
73	G	A
89	C	T
146	C	T
207	A	G
263	G	A
315.1	C	:
489	C	T
750	G	A
861	**C**	**T**
1,375	T	C
1,438	G	A
2,706	G	A
3,107	:	C
4,117	C	T
4,769	G	A
5,460	A	G
5,843	G	A
7,028	T	C
7,993	C	T
8,701	G	A
8,790	A	G
8,860	G	A
8,964	T	C
9,540	C	T
10,256	C	T
10,398	G	A
10,400	T	C
10,873	C	T
11,314	G	A
11,719	A	G
12,705	T	C
12,940	A	G
13,500	C	T
14,025	C	T
14,766	T	C
14,783	C	T
15,043	A	G
15,301	A	G
15,326	G	A
16,129	A	G
16,144	C	T
16,148	T	C
16,172	C	T
16,265	C	A
16,311	C	T
16,343	G	A
16,519	C	T

DQ372883 T726

43 differences to rCRS. Haplogroup M/M28

Base	T726	rCRS
73	G	A
152	C	T
195	C	T
263	G	A
315.1	C	:
489	C	T
750	G	A
1,438	G	A
1,598	A	G
1,719	A	G
2,706	G	A
3,107	:	C
4,769	G	A
6,281	G	A
6,374	C	T
6,858	**G**	**A**
6,962	A	G
7,028	T	C
8,701	G	A
8,860	G	A
9,540	C	T
10,192	**T**	**C**
10,245	C	T
10,398	G	A
10,400	T	C
10,658	G	A
10,873	C	T
11,719	A	G
12,705	T	C
14,766	T	C
14,783	C	T
15,043	A	G
15,067	C	T
15,301	A	G
15,326	G	A
15,530	**C**	**T**
16,086	C	T
16,129	A	G
16,148	T	C
16,223	T	C
16,311	**C**	**T**
16,362	C	T
16,519	C	T

DQ372884 CI153

54 differences to rCRS. Haplogroup M/Q1

Base	CI153	rCRS
73	G	A
89	C	T
92	A	G
146	C	T
263	G	A
309.1	C	:
315.1	C	:
489	C	T
514	:	C
515	:	A
750	G	A
1,407	A	T
1,438	G	A
2,706	G	A
3,107	:	C
3,834	A	G
4,117	C	T
4,769	G	A
4,913	C	A
5,460	A	G
5,843	G	A
7,028	T	C
8,701	G	A
8,790	A	G
8,860	G	A
8,964	T	C
9,101	C	T
9,540	C	T
10,398	G	A
10,400	T	C
10,873	C	T
11,719	A	G
11,884	G	A
12,705	T	C
12,940	A	G
13,047	G	A
13,500	C	T
14,025	C	T
14,766	T	C
14,783	C	T
14,798	C	T
15,043	A	G
15,301	A	G
15,326	G	A
16,129	A	G
16,144	C	T
16,148	T	C
16,223	T	C
16,241	G	A
16,265	C	A
16,293	**G**	**A**
16,311	C	T
16,343	G	A
16,526	A	G

DQ372885 WS72

54 differences to rCRS. Haplogroup M/Q1

Base	WS72	rCRS
73	G	A
89	C	T
92	A	G
146	C	T
263	G	A
309.1	C	:
309.2	C	:
315.1	C	:

489	C	T
514	:	C
515	:	A
750	G	A
1,407	A	T
1,438	G	A
2,706	G	A
3,107	:	C
3,834	A	G
4,117	C	T
4,769	G	A
4,913	C	A
5,460	A	G
5,843	G	A
7,028	T	C
8,701	G	A
8,790	A	G
8,860	G	A
8,964	T	C
9,101	C	T
9,540	C	T
10,398	G	A
10,400	T	C
10,873	C	T
11,719	A	G
11,884	G	A
12,705	T	C
12,940	A	G
13,047	G	A
13,500	C	T
14,025	C	T
14,766	T	C
14,783	C	T
14,798	C	T
15,043	A	G
15,301	A	G
15,326	G	A
16,129	A	G
16,144	C	T
16,148	T	C
16,223	T	C
16,241	G	A
16,265	C	A
16,311	C	T
16,343	G	A
16,526	A	G

DQ372886 TL36
42 differences to rCRS. Haplogroup N/R/B4a1a1/Pol. motif

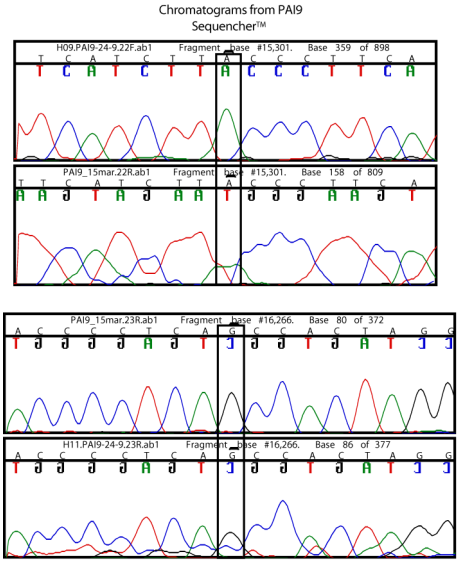
Base	TL36	Seq2
73	G	A
146	C	T
263	G	A
309.1	C	:
309.2	C	:
315.1	C	:
514	:	C
515	:	A
750	G	A
1,438	G	A
2,706	G	A
3,107	:	C
4,769	G	A
5,465	C	T
5,563	**A**	**G**
6,719	C	T
7,028	T	C
8,271	:	A
8,272	:	C
8,273	:	C
8,274	:	C
8,275	:	C
8,276	:	C
8,277	:	T
8,278	:	C
8,279	:	T
8,860	G	A
9,123	A	G
10,238	C	T
11,719	A	G
12,239	T	C
14,022	G	A
14,766	T	C
15,326	G	A
15,746	G	A
16,182	C	A
16,183	C	A
16,189	C	T
16,217	C	T
16,247	G	A
16,261	T	C
16,519	C	T

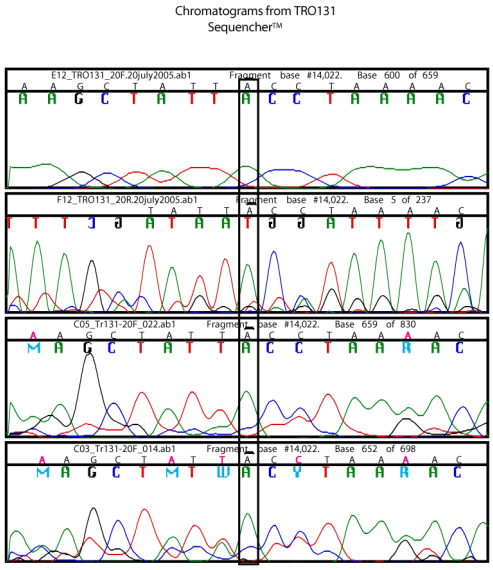
DQ372887 TRI65
42 differences to rCRS. Haplogroup N/W

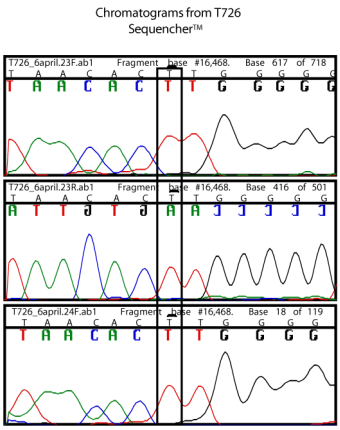
Base	TRI65	rCRS
73	G	A
189	G	A
194	T	C
195	C	T
199	C	T
204	C	T
207	A	G
263	G	A
309.1	C	:
309.2	C	:
315.1	C	:
709	A	G
750	G	A
1,243	C	T
1,406	C	T
1,438	G	A
2,706	G	A
3,107	:	C
3,505	G	A
4,769	G	A
5,046	A	G
5,460	A	G
7,028	T	C
7,874	G	A
8,251	A	G
8,860	G	A
8,994	A	G
10,398	G	A
11,674	T	C
11,719	A	G
11,947	G	A
12,414	C	T
12,705	T	C
13,263	G	A
13,344	G	A
14,766	T	C
15,326	G	A
15,784	C	T
15,884	C	G
16,223	T	C
16,292	T	C
16,519	C	T

Notes on specific polymorphisms

DQ372868: 9 bp deletion (AMI15, M7c). The presence of the 9bp deletion, from nucleotides 8280-8288, which defines haplogroups N/R/B4 and N/RB5 in this sequence is unusual, and unique to AMI15 in the M7c cluster. Five sequencing reads clearly identify the deletion, and two







of these extend as far as base 8701 which is a guanine in this sample as it generally is in all but macrohaplogroup N samples, where there is a transition to an adenine at the N vertex, acting as strong evidence against the accidental sequencing of a N/R/B sample in this region.

DQ372869: 15301A, 16266G (PAI9, B5a). This sequence has two unusual features in the B5a phylogeny. It has a transition at 15301 from G to A, and requires a transition at 16266 from A to G (a transversion at 16266 from C to A appears to be ancestral to the B5a cluster). As 15301A-G is one of the polymorphisms defining the N branch of the world tree this implies a reversion at the site in DQ372869. Both forward and reverse sequences confirm the 15301A base, and the presence of the B5a defining polymorphism 15235G in the same sequencing reads provides a strong argument for the validity of the 15301A reversion (as opposed to mistaken amplification and sequencing of a M-type template). The 16266G polymorphism is clearly shown on forward and reverse sequences.

DQ372871: 14022A. This sequence (TRO131) is interesting in that it is internal to the B4a1a cluster: six lineages descend from this node; five to Taiwanese sequences and one to all B4 sequences from Oceania cluster B4a1a1. Thus this sample is the only one from Oceania to date which does not belong to the B4a1a1 cluster but appears ancestral to it.

DQ372883: 16468T (T726, M28). A transition at 16468 from T to C is placed as ancestral to the M28 cluster in the labelled phylogeny. In DQ372883 the base at this position is a T, and thus requires a reversion at the site. Three sequencing reads (two forward, one reverse) clearly show 16468T. These sequencing reads encompass the region from 16050 - 16569, and show the other expected polymorphisms for haplogroup M28a (16086C, 16129A, 16148T, 16223T, 16362C).

E5.1 Character scores for random 75-taxon minimal tree sets (page 1/9)

rCRS number	States	Ave. steps all data sets (n=114)	% of datasets pars. inform.	Ave. steps when inform.
10	CT	0.1	8	1.0
41	CT	0.0	4	1.0
55	ACT	0.1	14	1.1
56	AT	0.2	17	1.0
57	ACGT	0.2	16	1.1
59	CT	0.0	2	1.0
61	CT	0.1	6	1.0
62	AGT	0.1	7	1.0
63	CT	0.3	22	1.2
64	CT	1.7	87	1.9
66	AGT	0.3	25	1.1
67	GT	0.1	12	1.0
68	AG	0.1	12	1.0
72	CT	1.0	78	1.2
73	ACG	2.8	100	2.8
75	AG	0.1	14	1.1
89	CT	0.2	23	1.0
92	AG	0.1	9	1.0
93	AG	3.0	100	3.0
94	AG	0.5	39	1.2
95	AC	1.5	93	1.7
103	ACG	0.4	30	1.2
114	CT	0.3	27	1.1
119	CT	0.0	4	1.0
125	CT	0.1	12	1.0
127	CT	0.2	15	1.0
128	CT	0.1	9	1.0
131	CT	0.3	27	1.0
143	AG	2.4	95	2.6
146	ACT	8.2	100	8.2
150	CGT	7.2	100	7.2
151	CT	3.8	100	3.8
152	CT	12.0	100	12.0
153	AG	0.7	53	1.3
154	CT	0.0	4	1.0
182	CT	3.2	100	3.2
183	AG	1.1	74	1.5
185	AGT	3.4	100	3.4
186	ACT	1.1	100	1.1
188	AG	0.4	40	1.1
189	ACG	4.7	100	4.7
192	ACT	0.4	41	1.0
193	AGT	0.4	39	1.0
194	ACT	1.5	82	1.8
195	ACT	8.5	100	8.5
196	CT	0.1	5	1.0
198	CT	3.8	99	3.8
199	CT	3.1	99	3.1
200	AG	1.6	87	1.9
202	AG	0.3	32	1.0
203	AG	0.3	30	1.1
204	CT	4.6	100	4.6
205	AG	0.1	9	1.0
207	AG	3.7	99	3.7
208	CT	0.0	3	1.0
210	AG	0.3	21	1.2
211	AG	0.1	11	1.2
212	CT	0.1	11	1.0
214	AG	0.3	29	1.2
215	AG	0.4	29	1.2
217	CT	0.5	39	1.3
225	AG	0.1	12	1.0
227	AGT	0.6	46	1.2
228	AGT	1.2	82	1.4
234	AG	0.3	29	1.2
235	AG	0.9	85	1.1
236	CT	1.7	100	1.7
239	CT	0.1	11	1.0
241	AG	0.0	3	1.0
242	CT	0.3	27	1.0
246	CT	0.2	20	1.0
247	AG	1.0	100	1.0
250	CT	0.4	36	1.0
252	CT	0.2	16	1.1
257	AG	0.1	10	1.0
260	AG	0.3	28	1.0
263	AG	0.8	64	1.2
264	CT	0.5	53	1.0
270	AG	0.0	4	1.0
271	CGT	0.2	19	1.2
279	CT	0.1	10	1.0
280	CGT	0.1	13	1.1
281	AG	0.1	5	1.0
282	CT	0.1	7	1.0
284	AG	0.0	4	1.0
285	CT	0.1	13	1.0
292	AT	0.5	48	1.0
295	ACT	1.0	86	1.2
297	AG	1.1	100	1.1
298	CT	0.1	15	1.0
299	AC	0.1	10	1.0
302	AC	0.2	16	1.0
310	CT	0.4	35	1.2
311	CT	0.3	25	1.0
316	ACG	2.5	99	2.6
318	CT	0.1	11	1.0
319	CT	0.1	10	1.0
320	CT	0.1	7	1.0
322	AG	0.1	7	1.0
325	CT	0.9	84	1.1
326	AG	0.1	10	1.0
334	CT	0.1	14	1.1
338	CT	0.2	17	1.1
340	CT	0.2	19	1.0
357	ACG	1.3	100	1.3
368	AG	0.1	7	1.0
373	AG	0.2	21	1.0
385	AGT	0.5	47	1.1
408	AT	0.2	18	1.1
418	CT	0.8	82	1.0
431	ACT	0.5	45	1.0
437	CT	0.2	17	1.1
447	CG	0.2	24	1.0
453	CT	0.1	6	1.0
456	CT	1.4	87	1.6
458	CT	0.1	8	1.0
460	CT	0.1	9	1.0
461	CT	0.1	8	1.0
462	CT	0.9	86	1.0
463	CT	0.0	4	1.0
464	AG	0.0	4	1.0
467	CT	0.7	73	1.0
471	CT	0.1	5	1.0
477	CGT	0.3	28	1.2
480	CT	0.1	14	1.0
481	CT	0.1	5	1.0
482	CT	0.9	61	1.5
485	CT	0.2	20	1.2
489	CT	1.9	100	1.9
493	AG	0.1	8	1.0
497	CT	0.8	69	1.1
499	AG	0.6	56	1.1
508	AG	0.4	37	1.0
509	CT	0.1	14	1.0
511	CT	0.1	6	1.0
512	AC	0.0	4	1.0
513	AG	1.2	80	1.5
537	CT	0.1	7	1.0
545	AG	0.0	4	1.0
546	AG	0.1	12	1.0
548	CT	0.1	5	1.0
569	CT	0.1	10	1.0
591	AC	0.1	5	1.0
593	ACT	0.5	42	1.2
629	CT	0.3	29	1.1
633	AG	0.4	34	1.0
654	CT	0.1	6	1.0
656	CT	0.1	8	1.0
663	AG	0.9	86	1.0
680	CT	0.8	82	1.0
681	CT	0.3	32	1.0
699	AG	0.1	7	1.0
709	AG	6.6	100	6.6
710	CT	1.1	100	1.1
721	CT	0.1	9	1.0
723	ACG	0.4	35	1.1
735	AG	0.1	8	1.0
738	AG	0.1	5	1.0
739	CT	0.1	9	1.0
740	AG	0.1	11	1.0
742	CT	0.0	4	1.0
750	AG	0.3	27	1.1
752	CT	0.5	48	1.0
761	AG	0.1	5	1.0
769	AG	1.0	100	1.0
789	CT	0.1	7	1.0
793	CGT	0.1	13	1.1
794	CT	0.1	5	1.0
825	AT	1.0	100	1.0
827	AG	0.7	63	1.2
850	CT	0.5	47	1.0
856	AG	0.2	21	1.0
870	CT	0.3	28	1.0
921	CT	0.2	17	1.1
930	AG	1.1	80	1.4
942	AG	0.0	4	1.0
951	AG	0.2	17	1.1
962	CT	3.9	100	3.9
978	AG	0.3	25	1.1
980	CT	0.3	32	1.1
983	CT	0.1	5	1.0
1005	CT	0.2	15	1.1
1007	ACG	0.4	39	1.0
1018	AG	1.0	100	1.0
1041	AG	0.3	29	1.0
1048	CT	1.4	100	1.4
1095	CT	0.1	13	1.0
1107	CT	0.6	63	1.0
1119	CT	0.6	54	1.0
1148	AG	0.1	9	1.0
1185	CT	0.1	5	1.2
1189	CT	0.8	81	1.0
1193	CT	0.0	4	1.0
1211	AG	0.3	32	1.1
1243	CT	1.0	75	1.3
1291	CT	0.3	26	1.0
1299	AG	0.0	4	1.0
1309	AG	0.0	4	1.0
1310	CT	0.4	36	1.1
1342	CT	0.1	8	1.0
1375	CT	0.1	12	1.0
1382	AC	0.9	88	1.0
1391	CT	0.1	12	1.0
1393	AG	0.1	9	1.0
1406	CT	0.4	34	1.1
1407	AT	0.0	4	1.0
1420	CT	0.7	63	1.1
1438	AG	2.4	100	2.4
1440	AG	0.0	4	1.0
1442	AG	1.4	99	1.4
1452	CT	0.0	4	1.0
1453	AG	0.1	5	1.0
1462	AGT	0.2	20	1.1
1503	AG	0.1	6	1.0
1508	CT	0.0	4	1.0
1520	CT	0.1	11	1.0
1524	AG	0.1	15	1.0
1536	AG	0.1	8	1.0
1541	CT	0.1	10	1.0
1555	AG	0.3	23	1.1
1598	AG	1.3	83	1.5
1625	AG	0.1	5	1.2
1664	AG	0.5	39	1.2
1692	AGT	0.2	18	1.2
1703	ACGT	0.1	14	1.0
1706	CT	0.9	88	1.0
1709	AG	0.5	39	1.2
1715	CT	0.1	12	1.0

E5.1 cont. Character scores for random 75-taxon minimal tree sets (page 2/9)

rCRS number	States	Ave. steps all data sets (n=114)	% of datasets pars. inform.	Ave. steps when inform.
1719	AG	1.6	87	1.8
1721	CT	0.1	14	1.0
1734	CT	0.1	6	1.0
1736	AG	0.9	86	1.0
1738	CT	1.0	100	1.0
1766	CT	0.3	32	1.0
1780	CT	0.3	33	1.0
1808	AG	0.1	11	1.0
1811	AG	1.3	95	1.4
1819	CT	0.1	6	1.0
1822	CT	0.1	13	1.0
1824	CT	0.1	12	1.0
1834	CT	0.1	12	1.0
1842	AG	0.1	8	1.0
1850	CT	0.1	9	1.1
1888	AG	1.3	89	1.5
1900	AG	0.1	6	1.0
1977	CT	0.1	15	1.0
2000	CT	0.4	35	1.2
2045	AG	0.1	15	1.0
2056	AG	0.1	14	1.0
2083	CT	0.5	39	1.2
2092	CT	0.0	4	1.0
2109	AG	0.1	8	1.0
2157	ACT	0.5	49	1.1
2158	CT	0.3	26	1.0
2218	CT	0.1	14	1.0
2220	AG	0.1	7	1.0
2225	AC	0.0	2	1.5
2245	ACG	1.4	97	1.4
2246	AG	0.2	17	1.0
2259	CT	0.2	23	1.0
2263	AC	0.1	6	1.0
2283	CT	0.3	25	1.1
2308	AG	0.8	75	1.0
2315	AG	0.1	8	1.0
2330	CT	0.1	7	1.0
2332	CT	1.1	99	1.1
2352	CT	1.6	100	1.6
2355	AG	0.0	4	1.0
2358	AG	0.9	88	1.0
2361	AG	0.1	11	1.1
2380	CT	0.2	20	1.1
2387	CT	0.1	11	1.0
2392	CT	0.1	7	1.0
2404	CT	0.1	7	1.0
2416	CT	1.1	100	1.1
2442	CT	0.1	15	1.0
2483	CT	0.2	18	1.0
2581	AG	0.1	9	1.0
2626	CT	0.8	74	1.1
2639	CT	0.1	14	1.0
2706	AGT	1.4	100	1.4
2755	AG	0.7	66	1.1
2758	AG	1.0	100	1.0
2763	CT	0.0	4	1.0
2766	CT	0.5	47	1.0
2768	AG	1.1	100	1.1
2772	CT	1.1	80	1.4
2789	CT	1.2	100	1.2
2831	AGT	0.4	36	1.1
2833	AG	0.1	7	1.0
2835	CT	0.2	23	1.0
2836	AC	0.5	47	1.0
2850	CT	0.1	14	1.0
2851	AG	0.0	4	1.0
2863	CT	0.7	64	1.0
2885	CT	1.4	100	1.4
2887	CT	0.1	6	1.0
3010	AG	3.9	100	3.9
3027	CT	0.0	4	1.0
3083	CT	0.1	13	1.0
3116	CT	0.2	19	1.0
3145	AG	0.1	12	1.0
3197	CT	0.8	81	1.0
3200	ACT	0.9	83	1.1
3202	CT	0.1	6	1.1
3203	AG	0.2	19	1.0
3204	CT	0.2	17	1.1
3206	CT	0.7	66	1.0
3221	AG	0.4	39	1.2
3278	CT	0.1	7	1.0
3290	CT	0.2	16	1.1
3306	CT	0.1	7	1.0
3308	CT	1.5	100	1.5
3316	AG	0.6	52	1.2
3336	CT	0.5	41	1.1
3337	AG	0.1	9	1.0
3338	CT	0.5	43	1.1
3348	AG	0.5	49	1.0
3372	CT	0.6	52	1.1
3384	AG	0.1	10	1.0
3391	AG	0.4	36	1.1
3394	CT	0.8	66	1.3
3395	AG	0.1	9	1.0
3396	CT	0.7	57	1.2
3397	AG	0.5	43	1.2
3398	CT	0.2	15	1.1
3421	AG	0.4	38	1.1
3434	AG	0.6	48	1.2
3438	AG	0.6	46	1.3
3447	AG	0.1	9	1.0
3450	CT	0.3	25	1.0
3480	AG	0.9	85	1.0
3483	AG	0.0	4	1.0
3486	CT	0.0	3	1.0
3495	AC	0.7	66	1.0
3496	GT	0.1	14	1.0
3497	CT	0.5	49	1.0
3504	CT	0.1	7	1.0
3505	AG	0.5	53	1.0
3510	CT	0.1	6	1.0
3513	CT	0.6	61	1.0
3516	AC	1.0	100	1.0
3531	AG	0.3	28	1.2
3535	CT	0.1	11	1.0
3537	AG	0.7	56	1.2
3540	ACT	0.3	24	1.1
3546	CT	0.1	8	1.0
3547	AG	0.1	8	1.0
3549	CT	0.1	11	1.0
3552	ACT	0.6	54	1.1
3591	AG	0.2	19	1.2
3593	CT	0.0	2	1.0
3594	CT	1.0	100	1.0
3606	AG	0.1	6	1.0
3630	CT	0.1	5	1.0
3644	CGT	0.4	37	1.2
3645	CT	0.1	13	1.0
3666	AG	1.1	100	1.1
3687	CT	0.1	6	1.0
3693	AG	1.4	100	1.4
3696	CT	0.1	10	1.1
3699	CT	0.0	4	1.0
3705	AG	0.3	28	1.0
3714	AG	0.1	8	1.0
3720	AG	0.3	32	1.0
3729	AG	0.1	6	1.0
3736	AG	0.3	27	1.1
3738	CT	0.1	10	1.0
3741	CT	0.1	14	1.0
3744	AG	0.1	10	1.2
3756	AG	0.2	24	1.0
3759	AG	0.2	17	1.0
3777	CT	1.0	73	1.3
3786	CT	0.1	10	1.1
3796	AGT	1.0	97	1.0
3816	AG	0.3	25	1.1
3817	CT	0.1	8	1.0
3826	CT	0.1	11	1.0
3834	AG	0.5	44	1.1
3843	AG	1.1	93	1.1
3847	CT	0.1	7	1.0
3849	AG	0.1	8	1.0
3852	CT	0.2	24	1.0
3866	CT	0.6	61	1.0
3882	AG	0.3	26	1.0
3915	AG	0.3	25	1.1
3918	AG	1.1	100	1.1
3921	ACT	0.3	30	1.1
3927	AG	0.6	61	1.0
3960	CT	0.0	4	1.0
3970	CT	0.8	82	1.0
3975	CT	0.2	18	1.0
3981	AG	0.3	30	1.1
3990	CT	0.2	16	1.0
3992	CT	0.2	23	1.0
3999	CT	0.2	17	1.0
4012	AG	0.1	5	1.0
4017	CT	0.1	7	1.1
4024	AG	0.2	20	1.0
4025	CT	0.4	37	1.1
4047	CT	0.1	15	1.0
4048	AG	0.8	72	1.1
4050	CT	0.0	4	1.0
4059	CT	0.1	13	1.0
4062	CT	0.2	18	1.0
4071	CT	0.8	79	1.0
4080	CT	0.1	7	1.0
4086	CT	0.7	59	1.2
4092	AG	0.1	7	1.0
4093	AG	0.1	8	1.1
4099	CT	0.1	11	1.0
4104	AG	1.3	100	1.3
4113	AG	0.1	10	1.1
4117	CT	0.6	55	1.1
4122	AG	0.2	18	1.0
4129	AG	0.1	6	1.0
4140	CT	0.1	15	1.0
4158	AG	0.9	88	1.0
4161	CT	0.0	3	1.0
4164	AG	0.7	72	1.0
4181	AG	0.1	8	1.1
4185	CT	0.5	53	1.0
4188	AG	0.1	7	1.0
4200	AGT	0.1	8	1.0
4203	AG	0.1	8	1.0
4209	CT	0.1	13	1.0
4216	CT	1.5	100	1.5
4218	CT	0.2	17	1.1
4225	AG	0.3	31	1.1
4227	AG	0.1	5	1.0
4230	CT	0.0	4	1.0
4232	CT	0.4	32	1.1
4248	CT	0.9	87	1.1
4257	AG	0.1	6	1.0
4310	AG	0.1	11	1.1
4312	CT	1.0	100	1.0
4335	CT	0.1	7	1.0
4336	CT	0.4	44	1.0
4342	AG	0.1	8	1.0
4343	AG	0.4	44	1.0
4370	CT	0.9	88	1.0
4386	CT	0.8	76	1.1
4394	CT	0.1	12	1.0
4395	AG	0.1	12	1.0
4435	AG	0.3	33	1.0
4452	CT	0.1	12	1.0
4454	ACT	1.2	85	1.4
4491	AG	0.4	39	1.0
4505	CT	0.3	25	1.0
4506	AG	0.6	61	1.0
4508	CT	0.1	9	1.0
4529	AGT	0.4	40	1.0
4538	CT	0.1	11	1.0
4541	AGT	0.7	56	1.2
4550	CT	0.0	4	1.0
4561	CT	0.7	58	1.3

E5.1 cont. Character scores for random 75-taxon minimal tree sets (page 3/9)

rCRS number	States	Ave. steps all data sets (n=114)	% of datasets pars. inform.	Ave. steps when inform.
4562	AG	0.7	54	1.2
4580	AG	0.8	78	1.1
4586	CT	1.1	100	1.1
4612	CT	0.1	9	1.0
4639	CT	0.4	39	1.1
4640	ACT	0.1	9	1.1
4646	CT	0.2	21	1.0
4655	AG	1.2	82	1.5
4659	AG	0.1	11	1.2
4670	CT	0.2	17	1.0
4674	AG	0.0	4	1.0
4679	CT	0.2	15	1.1
4688	ACT	0.7	51	1.4
4695	CT	0.1	12	1.1
4703	CT	0.2	17	1.0
4705	CT	0.2	18	1.0
4715	AG	0.9	82	1.1
4732	AG	0.6	50	1.1
4733	CT	0.1	13	1.1
4735	AC	0.1	7	1.0
4736	CT	0.1	8	1.0
4742	CT	0.4	35	1.1
4745	AG	0.2	23	1.0
4767	AG	0.9	88	1.0
4769	AG	0.3	31	1.1
4775	AG	0.1	8	1.0
4790	AG	0.0	4	1.0
4793	AG	0.7	58	1.2
4808	CT	0.0	3	1.0
4820	AG	0.8	60	1.3
4823	CT	0.1	11	1.0
4824	AG	1.2	91	1.3
4833	AG	0.9	83	1.0
4850	CT	0.2	21	1.0
4859	CT	0.4	36	1.0
4883	CT	1.0	100	1.0
4895	AG	0.3	27	1.0
4907	CT	0.7	56	1.3
4913	AC	0.0	4	1.0
4916	AG	0.1	13	1.0
4917	AG	1.0	86	1.2
4928	CT	0.3	25	1.1
4958	AG	0.8	76	1.0
4959	AG	0.1	9	1.0
4960	CT	0.1	14	1.0
4965	AG	0.1	9	1.0
4976	AG	0.1	9	1.0
4977	CT	0.1	9	1.1
4991	AG	0.2	20	1.2
5004	CT	0.3	24	1.1
5021	CT	0.1	7	1.0
5027	CT	0.9	88	1.0
5036	AG	1.0	100	1.0
5046	AG	1.6	100	1.6
5048	CT	0.5	41	1.1
5049	CT	0.3	32	1.0
5051	AG	0.0	4	1.0
5063	CT	0.1	6	1.0
5081	CT	0.1	9	1.0
5082	ACT	0.1	8	1.0
5086	CT	0.0	4	1.3
5093	CT	0.2	18	1.1
5096	CT	0.9	88	1.0
5102	AG	0.1	6	1.0
5108	CT	1.3	94	1.4
5120	AG	0.1	9	1.0
5127	AG	0.1	9	1.0
5147	AG	2.4	95	2.6
5153	AG	0.7	52	1.3
5177	AG	0.1	15	1.0
5178	ACT	1.1	100	1.1
5186	AT	0.1	12	1.0
5198	AG	0.1	9	1.0
5201	CT	0.4	36	1.1
5206	CT	0.1	7	1.0
5213	CT	0.0	3	1.0
5231	AG	1.7	99	1.8
5237	AG	1.7	90	1.9
5240	AG	0.1	10	1.0
5250	CT	0.1	7	1.0
5252	AG	0.3	29	1.1
5261	AG	0.1	7	1.0
5262	AG	0.1	13	1.0
5263	CT	0.5	42	1.3
5268	AG	0.1	9	1.0
5276	AG	0.1	5	1.0
5285	AG	1.0	100	1.0
5298	AG	0.1	7	1.0
5300	CT	0.3	34	1.0
5301	AG	0.7	68	1.1
5302	CT	0.1	7	1.0
5319	AG	0.1	14	1.0
5324	CT	0.1	9	1.0
5331	AC	0.9	88	1.0
5348	CT	0.1	5	1.0
5351	AG	0.8	75	1.0
5360	CT	0.2	20	1.0
5375	CT	0.2	22	1.0
5378	AG	0.1	6	1.0
5379	CT	0.1	11	1.0
5390	AG	0.3	31	1.0
5393	CT	1.0	100	1.0
5417	AG	1.0	87	1.1
5423	AG	0.1	10	1.0
5426	CT	0.4	37	1.1
5432	AG	0.2	18	1.0
5437	CT	0.0	3	1.0
5441	AGT	0.1	14	1.0
5442	CT	1.3	100	1.3
5452	CT	0.1	10	1.0
5460	AG	3.8	100	3.8
5465	CT	0.7	73	1.0
5466	AG	0.1	9	1.0
5471	AG	0.6	48	1.3
5478	CT	0.0	4	1.0
5483	CT	0.3	32	1.0
5486	CT	0.0	4	1.0
5492	CT	0.2	21	1.1
5493	CT	0.1	11	1.0
5495	CT	0.3	32	1.0
5498	AG	0.1	7	1.0
5503	CT	0.0	4	1.0
5510	AG	0.1	7	1.0
5539	AG	0.1	6	1.0
5557	CT	0.1	8	1.0
5558	AG	0.1	10	1.0
5563	AG	0.4	35	1.1
5580	CT	0.4	39	1.1
5581	AG	0.9	89	1.0
5582	AG	0.1	8	1.0
5585	AG	0.3	25	1.1
5592	AG	0.0	3	1.0
5601	CT	0.8	71	1.1
5603	CT	1.0	97	1.0
5628	CT	0.4	32	1.1
5633	CT	0.0	4	1.0
5655	CT	1.3	100	1.3
5656	AG	0.6	58	1.0
5711	AG	0.6	60	1.0
5744	AG	0.8	64	1.2
5747	AG	0.1	8	1.0
5772	AG	0.0	4	1.0
5773	AG	1.1	75	1.4
5774	CT	0.1	10	1.1
5783	AG	0.0	4	1.0
5788	CT	0.1	10	1.1
5790	AC	0.1	5	1.0
5811	AG	0.5	47	1.0
5814	CT	0.9	89	1.0
5821	AG	0.2	14	1.1
5824	AG	0.1	12	1.0
5826	CT	0.4	43	1.0
5836	AG	0.1	11	1.0
5843	AG	0.3	32	1.1
5846	CT	0.0	4	1.0
5892	CT	0.1	11	1.0
5893	CT	0.0	3	1.0
5894	ACGT	0.3	22	1.3
5911	CT	0.8	73	1.1
5913	AG	0.2	18	1.0
5922	CT	0.1	7	1.3
5951	AG	1.1	100	1.1
5964	CT	0.2	20	1.0
5972	CT	0.1	11	1.0
5978	AG	0.1	15	1.0
5984	AG	0.8	74	1.1
5987	CT	0.1	6	1.0
5999	CT	0.2	18	1.1
6005	CT	0.0	3	1.0
6018	AG	0.1	6	1.0
6023	AG	0.6	51	1.2
6026	AG	1.2	91	1.3
6032	AG	0.1	7	1.0
6040	AG	0.1	12	1.0
6045	CT	0.3	28	1.0
6047	AG	0.2	16	1.0
6053	CT	0.1	8	1.0
6071	CT	1.0	100	1.0
6077	CT	0.2	16	1.0
6086	CT	0.1	9	1.0
6104	CT	0.0	4	1.0
6125	AG	0.3	26	1.0
6131	AG	0.0	4	1.0
6146	AG	0.1	5	1.0
6150	AG	0.7	72	1.0
6152	CT	0.9	75	1.3
6164	CT	0.6	56	1.1
6167	CGT	0.2	22	1.0
6179	AG	0.4	39	1.0
6182	AG	0.6	57	1.1
6185	CT	1.2	100	1.2
6216	CT	0.3	24	1.1
6221	ACT	1.9	94	2.1
6227	CT	0.0	4	1.0
6248	CT	0.1	10	1.0
6249	AG	0.1	11	1.0
6253	CT	1.2	84	1.4
6257	AG	0.6	60	1.0
6260	AG	0.6	51	1.2
6261	AG	0.1	12	1.0
6263	CT	0.1	5	1.0
6267	AG	0.1	11	1.0
6281	AG	0.1	13	1.0
6293	CT	0.2	24	1.0
6297	CT	0.3	32	1.1
6320	CT	0.0	4	1.0
6332	AG	0.1	9	1.0
6345	CT	0.1	6	1.0
6351	CT	0.1	9	1.0
6353	AG	0.3	31	1.0
6359	AG	0.3	25	1.0
6365	CT	0.2	19	1.0
6366	AG	0.3	25	1.1
6371	CT	0.1	12	1.0
6374	CT	0.1	13	1.0
6378	CT	0.3	30	1.0
6383	AG	0.1	7	1.0
6386	CT	0.0	4	1.0
6392	CT	1.0	85	1.2
6410	CT	0.4	39	1.0
6413	CT	0.4	38	1.0
6437	AG	0.3	32	1.0
6446	AG	0.6	50	1.1
6455	CT	1.0	95	1.1
6473	CT	0.1	8	1.0
6485	AG	0.1	6	1.0
6515	CT	0.1	11	1.0
6518	CT	0.1	10	1.0

E5.1 cont. Character scores for random 75-taxon minimal tree sets (page 4/9)

rCRS number	States	Ave. steps all data sets (n=114)	% of datasets pars. inform.	Ave. steps when inform.
6524	CT	0.4	35	1.1
6531	CT	0.1	10	1.0
6533	AG	0.1	5	1.0
6548	CT	1.0	100	1.0
6575	AG	0.1	10	1.3
6584	CT	0.1	5	1.0
6587	CT	0.2	22	1.0
6599	AG	0.1	6	1.0
6620	CT	0.3	30	1.1
6629	AG	0.5	44	1.1
6632	CT	0.1	9	1.1
6647	AG	0.1	5	1.0
6653	ACT	0.1	12	1.0
6663	AG	0.6	61	1.0
6671	CT	0.2	17	1.0
6680	CT	0.8	75	1.1
6689	CT	0.4	42	1.0
6713	CT	1.0	89	1.1
6719	CT	0.8	67	1.2
6734	AG	0.4	33	1.1
6737	AG	0.3	25	1.0
6752	AG	1.2	80	1.5
6755	AG	0.1	12	1.0
6764	AG	0.1	5	1.0
6770	AG	0.0	4	1.0
6776	CT	0.6	57	1.0
6779	AG	0.1	10	1.0
6794	AG	0.1	10	1.0
6806	AG	0.1	6	1.0
6815	CT	0.3	25	1.0
6827	CT	1.1	100	1.1
6881	AG	0.1	6	1.0
6899	AG	0.1	15	1.0
6905	AG	0.1	12	1.0
6908	CT	0.1	11	1.0
6917	AG	0.8	76	1.1
6929	AG	0.0	4	1.0
6938	CT	0.5	50	1.0
6941	CT	0.1	10	1.0
6956	CGT	0.2	17	1.1
6960	CT	0.2	16	1.0
6962	AG	0.9	76	1.2
6983	CT	0.0	4	1.0
6989	AG	1.0	100	1.0
7022	CT	0.1	14	1.0
7028	CT	1.0	100	1.0
7046	AG	0.0	4	1.0
7052	AG	0.1	6	1.0
7055	AG	1.8	100	1.8
7058	ACT	0.1	9	1.0
7076	AG	0.7	73	1.0
7082	CT	0.0	3	1.0
7091	AG	0.1	5	1.0
7094	CT	0.1	7	1.0
7142	CT	0.1	5	1.0
7146	AG	1.3	100	1.3
7148	CT	0.3	25	1.1
7158	AG	0.1	5	1.0
7175	CT	1.1	100	1.1
7184	AG	0.1	7	1.0
7193	CT	0.1	9	1.0
7196	ACT	0.8	81	1.0
7202	AG	0.7	61	1.1
7214	CT	0.1	11	1.0
7220	CT	0.1	11	1.2
7226	AG	0.1	9	1.1
7245	AG	0.1	9	1.1
7250	AG	0.1	15	1.0
7256	CT	1.1	100	1.1
7257	AG	0.5	47	1.0
7258	CT	0.3	30	1.1
7268	GT	0.1	10	1.0
7270	CT	0.2	19	1.1
7271	AG	0.1	15	1.0
7274	CT	1.0	100	1.0
7298	ACG	0.1	13	1.0
7309	CT	0.1	7	1.0
7319	CT	0.1	6	1.1
7337	AG	0.9	79	1.2
7340	AG	0.1	8	1.1
7364	AG	0.1	5	1.0
7385	AG	0.3	31	1.0
7388	AG	0.0	1	1.0
7389	CT	1.1	100	1.1
7391	CT	0.3	28	1.1
7403	AG	0.2	16	1.1
7424	AG	0.2	21	1.0
7444	AG	0.2	16	1.1
7476	CT	0.1	11	1.0
7490	AG	0.1	10	1.0
7492	CT	0.1	6	1.0
7493	CT	0.1	5	1.0
7498	AG	0.6	55	1.1
7521	AG	1.3	100	1.3
7533	CT	0.1	8	1.1
7559	AG	0.1	6	1.0
7569	AG	0.1	8	1.0
7581	CT	0.1	11	1.0
7598	AG	0.2	16	1.0
7600	AG	0.7	67	1.0
7618	CT	0.1	9	1.0
7621	CT	0.1	9	1.0
7624	AT	1.0	98	1.0
7642	AG	0.1	10	1.0
7645	CT	0.1	7	1.0
7657	CT	0.0	4	1.0
7660	CT	0.5	50	1.0
7664	AG	0.4	35	1.1
7673	AG	0.4	35	1.2
7681	ACT	0.2	22	1.0
7684	CT	0.7	71	1.0
7693	CT	0.5	47	1.0
7697	AG	0.0	2	1.0
7702	AG	0.4	37	1.0
7705	CT	0.1	6	1.0
7711	CT	0.1	13	1.1
7738	CT	0.1	6	1.0
7759	CT	0.1	13	1.0
7762	AG	0.1	11	1.0
7765	AG	0.1	9	1.0
7768	AG	0.7	68	1.0
7771	AG	1.0	100	1.0
7785	CT	0.0	4	1.0
7789	AG	0.7	64	1.1
7805	AG	0.7	54	1.3
7819	ACT	0.1	12	1.0
7822	AG	0.2	18	1.0
7828	AG	0.1	12	1.0
7830	AG	0.1	11	1.0
7844	AG	0.3	31	1.1
7852	AG	0.1	8	1.0
7853	AG	0.8	72	1.1
7859	AG	0.1	12	1.0
7861	CT	0.3	20	1.3
7864	CT	0.5	53	1.0
7867	CT	1.5	100	1.5
7870	CT	0.0	3	1.0
7891	CT	0.1	7	1.0
7909	CT	0.1	11	1.0
7912	AG	0.1	10	1.2
7927	CGT	0.1	11	1.0
7948	CT	0.3	31	1.1
7961	CT	0.2	24	1.0
7993	CT	0.1	6	1.0
7999	CT	0.1	12	1.0
8005	CT	0.1	9	1.0
8014	AGT	0.1	13	1.0
8020	AG	1.2	96	1.3
8023	CT	0.1	9	1.0
8027	AG	1.2	100	1.2
8063	CT	0.1	12	1.0
8071	AG	0.1	13	1.0
8074	AG	0.1	6	1.0
8075	AG	0.1	10	1.0
8078	AG	0.1	8	1.0
8080	CT	0.9	88	1.0
8087	CT	0.9	92	1.0
8108	AG	0.1	11	1.1
8110	CT	0.0	1	1.0
8137	CT	0.1	14	1.0
8149	AG	0.2	22	1.1
8152	AG	0.2	22	1.1
8158	AG	0.1	10	1.0
8164	CT	0.1	8	1.0
8167	ACT	0.2	22	1.1
8188	AG	0.1	14	1.0
8191	AG	0.4	43	1.0
8200	CT	0.5	52	1.0
8206	AG	1.3	100	1.3
8227	CT	0.4	38	1.1
8248	AG	1.0	100	1.0
8251	AG	2.7	98	2.7
8260	CT	0.1	8	1.0
8269	AG	0.7	56	1.3
8270	CT	0.3	24	1.1
8271	AT	0.2	16	1.0
8277	CT	0.4	37	1.1
8279	CT	0.3	31	1.0
8291	AG	0.0	4	1.0
8292	AG	0.3	32	1.1
8296	AG	0.1	5	1.0
8308	AG	0.1	10	1.0
8347	AG	0.0	4	1.0
8376	CT	0.1	7	1.1
8383	CT	0.4	37	1.1
8387	AG	0.9	89	1.0
8389	AG	0.1	9	1.1
8392	AG	0.4	38	1.1
8393	CT	0.1	5	1.0
8396	AG	0.2	18	1.0
8400	CT	0.0	4	1.0
8404	CT	0.5	48	1.1
8406	CT	0.1	7	1.0
8410	CT	0.0	4	1.0
8413	AG	0.0	4	1.0
8414	CT	1.0	100	1.0
8419	CT	0.4	32	1.1
8428	CT	1.0	97	1.0
8433	CT	0.0	3	1.0
8435	AG	0.0	4	1.0
8440	AG	0.1	7	1.0
8448	CT	0.1	5	1.0
8450	CT	0.2	17	1.0
8453	AG	0.1	11	1.0
8460	AG	0.6	60	1.0
8468	CT	1.1	100	1.1
8470	AG	0.1	9	1.0
8472	CT	0.1	12	1.0
8473	CT	0.8	70	1.1
8485	AG	0.2	16	1.1
8502	AG	0.3	28	1.0
8503	CT	0.1	7	1.0
8506	CT	0.1	11	1.0
8521	AG	0.0	4	1.0
8527	AG	0.1	9	1.0
8530	AG	0.0	4	1.0
8537	AG	0.1	11	1.0
8541	AG	0.5	51	1.0
8545	AG	0.2	23	1.0
8551	CT	0.0	4	1.0
8553	CT	0.1	6	1.0
8557	AG	0.5	41	1.2
8562	CT	0.1	10	1.0
8563	AG	0.8	75	1.0
8566	AG	1.1	98	1.1
8567	CT	0.0	4	1.0
8572	AG	0.3	27	1.0
8573	AG	0.1	7	1.0

E5.1 cont. Character scores for random 75-taxon minimal tree sets (page 5/9)

rCRS number	States	Ave. steps all data sets (n=114)	% of datasets pars. inform.	Ave. steps when inform.
8575	CT	0.1	8	1.0
8577	AG	0.3	30	1.1
8584	AG	1.5	96	1.5
8592	AG	0.1	7	1.0
8594	CT	0.2	24	1.0
8595	CT	0.1	6	1.0
8603	CT	0.1	8	1.0
8604	CT	0.5	47	1.0
8614	CT	0.1	8	1.0
8616	AGT	0.2	22	1.1
8618	CT	0.2	22	1.0
8628	CT	0.1	9	1.0
8650	CT	0.1	13	1.0
8654	CT	0.1	7	1.1
8655	CT	1.0	100	1.0
8676	CT	0.1	5	1.0
8679	AG	0.2	19	1.0
8684	CT	0.5	46	1.1
8697	AG	0.9	83	1.1
8701	AG	1.6	100	1.6
8703	CT	0.1	10	1.0
8705	CT	0.4	34	1.2
8709	CT	0.1	5	1.0
8718	AG	0.1	6	1.0
8730	AG	0.1	5	1.0
8762	CT	0.4	39	1.0
8764	AG	0.7	57	1.2
8784	AGT	1.0	77	1.3
8790	AG	1.6	92	1.7
8793	CT	0.3	25	1.1
8794	CT	0.9	86	1.0
8805	AG	0.1	11	1.0
8812	AG	0.1	11	1.0
8818	CT	0.1	11	1.0
8829	CT	0.4	44	1.0
8836	AG	0.1	14	1.0
8838	AG	0.1	12	1.0
8839	AG	0.1	11	1.2
8842	ACG	0.1	5	1.0
8844	CT	0.1	6	1.0
8854	AG	0.1	7	1.0
8856	AG	0.8	61	1.2
8859	CT	0.1	15	1.0
8860	AG	0.0	4	1.0
8865	AG	0.0	4	1.0
8869	AG	0.3	29	1.0
8870	CT	0.1	7	1.1
8875	CT	0.1	6	1.0
8877	CT	0.7	72	1.0
8886	AG	0.1	5	1.0
8894	AT	0.1	12	1.0
8911	CT	0.5	51	1.1
8923	AG	0.1	12	1.0
8925	AG	0.6	60	1.1
8928	CT	0.6	56	1.0
8932	CT	0.1	11	1.0
8943	CT	0.1	10	1.0
8950	AG	0.1	8	1.0
8952	CT	0.1	6	1.0
8962	AG	0.1	8	1.0
8964	CT	1.0	88	1.2
8973	AG	0.2	18	1.0
8987	CT	0.3	28	1.1
8988	AG	0.0	4	1.0
8994	AG	1.1	79	1.4
9004	CT	0.1	8	1.0
9007	AG	0.1	13	1.0
9033	AG	0.1	11	1.0
9041	AG	0.1	11	1.0
9042	CT	1.0	100	1.0
9051	AG	0.1	10	1.0
9052	AG	0.4	34	1.1
9053	AG	0.7	56	1.2
9055	AG	0.9	85	1.0
9058	AG	0.1	5	1.2
9064	AG	0.2	16	1.1
9066	AG	0.1	7	1.0
9072	AG	1.0	100	1.0
9077	CT	0.1	13	1.0
9083	CT	0.0	4	1.0
9084	CT	0.1	6	1.0
9088	CT	0.1	11	1.0
9090	CT	0.4	38	1.0
9091	AG	0.1	9	1.0
9093	ACG	0.3	27	1.0
9094	CT	0.1	9	1.0
9101	CT	0.1	6	1.0
9103	CT	0.1	14	1.1
9109	AG	0.1	11	1.0
9110	CT	0.1	10	1.0
9111	CT	0.0	2	1.0
9115	AG	0.0	2	1.0
9116	CT	0.1	8	1.0
9120	AG	0.1	8	1.0
9123	AG	1.0	80	1.2
9127	AG	0.1	7	1.0
9128	CT	0.1	11	1.2
9129	CT	0.1	11	1.1
9136	AG	0.5	47	1.0
9137	CT	0.1	7	1.0
9139	AG	0.1	8	1.1
9140	CT	0.1	10	1.0
9145	AG	0.1	7	1.0
9148	CT	0.1	6	1.0
9150	AG	0.1	12	1.0
9156	AG	0.1	9	1.0
9165	CT	0.2	22	1.0
9174	CT	0.2	17	1.1
9180	AG	0.6	60	1.0
9181	AG	0.3	25	1.1
9182	AG	0.1	8	1.0
9201	CT	0.2	22	1.0
9221	AG	1.1	100	1.1
9242	AG	0.3	32	1.1
9254	AG	0.4	32	1.1
9263	AG	0.0	4	1.0
9266	AG	0.2	17	1.1
9272	CT	0.5	47	1.0
9288	AG	0.0	2	1.0
9296	CT	0.7	71	1.0
9299	AG	0.2	16	1.1
9300	AG	0.4	39	1.1
9305	AG	0.1	8	1.0
9311	CT	0.6	56	1.0
9325	CT	0.1	7	1.1
9329	AG	0.1	12	1.1
9335	CT	0.1	8	1.0
9336	ACG	0.4	32	1.1
9347	AG	1.0	100	1.0
9355	AG	0.4	36	1.1
9365	CT	0.1	7	1.0
9374	AG	0.0	4	1.0
9377	AG	1.4	82	1.7
9380	AG	0.2	16	1.0
9386	CT	0.1	6	1.0
9389	AG	0.2	18	1.0
9410	AG	0.1	7	1.0
9431	CT	0.1	5	1.0
9438	AG	0.2	23	1.1
9449	CT	0.4	36	1.2
9452	AG	0.0	2	1.0
9468	AG	0.1	9	1.0
9477	AG	0.9	78	1.2
9479	CT	0.1	7	1.0
9480	CT	0.1	7	1.0
9494	AG	0.2	18	1.0
9495	CT	0.1	7	1.0
9509	CT	0.0	3	1.0
9527	CT	0.1	6	1.0
9530	CT	0.1	6	1.0
9536	CT	0.2	23	1.0
9540	CT	1.0	100	1.0
9545	AGT	0.9	68	1.3
9548	AG	0.6	52	1.1
9554	AG	0.6	49	1.3
9575	AG	0.7	68	1.0
9581	CT	0.4	36	1.2
9591	AG	0.0	2	1.0
9612	AG	0.3	26	1.0
9614	AG	0.1	9	1.0
9617	AG	0.1	8	1.0
9632	AG	0.1	8	1.0
9635	AC	0.1	11	1.0
9647	CT	0.6	61	1.0
9653	CT	0.0	4	1.0
9656	CT	0.1	9	1.0
9667	AG	0.3	26	1.0
9670	AG	0.0	4	1.0
9682	CT	0.1	12	1.0
9698	CT	0.9	87	1.0
9716	CT	0.6	49	1.3
9738	AG	0.1	11	1.1
9750	CT	0.1	6	1.0
9755	AG	1.4	98	1.4
9758	CT	0.7	59	1.3
9767	CT	0.0	4	1.0
9804	AG	0.4	30	1.3
9812	CT	0.1	6	1.0
9818	CT	1.0	100	1.0
9822	AC	0.1	6	1.0
9824	ACT	2.1	98	2.1
9830	CT	0.1	5	1.0
9833	CT	0.1	10	1.0
9852	AG	0.0	4	1.0
9861	CT	0.3	27	1.0
9899	CT	0.4	41	1.1
9921	AG	0.0	4	1.0
9932	AG	0.5	40	1.2
9938	CT	0.2	15	1.1
9941	AG	0.3	28	1.0
9944	CT	0.1	8	1.0
9947	AG	0.2	18	1.0
9948	AG	0.0	4	1.0
9950	CT	0.8	68	1.2
9959	CT	0.0	4	1.0
9962	AG	0.2	19	1.1
9966	AG	0.7	62	1.2
9977	CT	0.1	11	1.0
9986	AG	0.5	45	1.2
10005	AG	0.1	5	1.0
10007	CT	0.1	7	1.1
10031	CT	0.4	35	1.0
10034	CT	0.6	46	1.2
10042	AG	0.1	9	1.0
10044	AG	0.2	23	1.0
10070	CT	0.0	4	1.0
10084	CT	0.9	64	1.5
10086	AG	0.4	36	1.2
10088	CT	0.1	5	1.0
10097	ACG	0.1	13	1.0
10104	CT	0.4	39	1.0
10115	CT	1.0	100	1.0
10118	CT	0.1	12	1.0
10142	CT	0.2	22	1.0
10143	AG	0.3	32	1.1
10166	CT	0.1	12	1.0
10172	AG	0.1	11	1.0
10181	CT	0.5	51	1.0
10188	AG	0.1	10	1.0
10192	ACT	0.1	8	1.0
10197	AG	0.0	4	1.0
10211	CT	0.1	12	1.0
10214	CT	0.1	8	1.0
10235	CT	0.1	5	1.0
10237	CT	0.1	6	1.0
10238	CT	1.2	81	1.4
10245	CT	0.1	15	1.0
10253	CT	0.1	4	1.2

E5.1 cont. Character scores for random 75-taxon minimal tree sets (page 6/9)

rCRS number	States	Ave. steps all data sets (n=114)	% of datasets pars. inform.	Ave. steps when inform.
10256	CT	0.1	8	1.0
10283	AG	0.1	8	1.0
10289	AG	0.2	18	1.1
10310	AG	1.0	89	1.2
10320	AG	0.1	6	1.0
10321	CT	1.0	100	1.0
10325	AG	0.2	19	1.0
10345	CT	0.7	67	1.0
10358	AG	0.1	8	1.1
10361	CT	0.1	11	1.0
10364	AG	0.0	4	1.0
10373	AG	0.5	42	1.2
10394	CT	0.1	13	1.0
10397	AG	0.7	64	1.0
10398	AG	4.2	100	4.2
10400	CT	1.0	100	1.0
10410	CT	0.7	65	1.1
10411	AG	0.1	11	1.0
10427	AG	0.1	7	1.0
10454	CT	0.8	62	1.3
10463	CT	0.9	84	1.1
10493	CT	0.1	8	1.0
10497	CT	0.1	9	1.0
10499	AG	0.5	50	1.1
10506	AG	0.1	10	1.0
10535	CT	0.1	12	1.0
10550	AG	0.8	76	1.1
10556	CT	0.1	5	1.0
10586	AG	1.1	100	1.1
10589	AG	1.1	100	1.1
10598	AG	0.0	4	1.0
10607	CT	0.5	47	1.0
10609	CT	0.7	71	1.0
10619	CT	0.1	11	1.0
10631	ACT	0.1	12	1.0
10637	CT	0.1	9	1.0
10640	CT	0.1	12	1.0
10646	AG	0.3	31	1.1
10658	AG	0.2	18	1.0
10664	CT	1.0	100	1.0
10667	CT	0.1	14	1.0
10670	CT	0.2	20	1.0
10685	AG	0.3	29	1.2
10688	AG	1.1	100	1.1
10700	AG	0.4	33	1.1
10724	CT	0.1	9	1.1
10733	CT	0.1	7	1.0
10736	CT	0.1	6	1.0
10750	AG	0.1	11	1.0
10754	ACG	0.4	35	1.1
10786	CT	0.1	10	1.0
10790	CT	0.9	67	1.3
10792	AG	0.7	72	1.0
10793	CT	0.7	72	1.0
10801	AG	0.6	57	1.0
10810	CT	1.2	100	1.2
10816	AGT	0.1	6	1.0
10819	AG	0.8	63	1.2
10828	CT	0.8	82	1.0
10873	CT	1.1	100	1.1
10876	AG	0.8	63	1.3
10914	AG	0.1	5	1.0
10915	CT	1.6	100	1.6
10920	CT	0.9	73	1.3
10927	CT	0.4	45	1.0
10939	CT	0.5	47	1.0
10940	CT	0.1	6	1.0
10972	AG	0.1	6	1.0
10976	CT	0.4	41	1.0
10978	AG	0.3	26	1.0
10986	AC	0.1	5	1.0
10993	AG	0.0	4	1.0
11002	AG	0.3	26	1.0
11009	CT	0.1	11	1.0
11013	CT	0.1	9	1.0
11014	CT	0.1	15	1.0
11016	AG	0.6	57	1.1
11017	CT	1.1	82	1.3
11020	AG	0.1	13	1.0
11025	CT	0.2	18	1.0
11038	AG	0.1	8	1.0
11047	AC	0.1	5	1.0
11050	CT	0.1	8	1.0
11059	CT	0.1	8	1.0
11061	CT	0.1	13	1.0
11065	AG	0.2	17	1.0
11077	AG	0.3	23	1.2
11083	AG	0.1	14	1.0
11084	AG	0.7	70	1.1
11092	AG	0.1	5	1.2
11101	AG	0.1	12	1.0
11113	CT	0.1	7	1.0
11143	CT	0.2	23	1.0
11146	CT	0.2	22	1.0
11147	CT	0.1	11	1.0
11149	AG	0.1	12	1.0
11150	AG	0.1	6	1.0
11151	CT	0.2	22	1.1
11167	AG	0.9	78	1.2
11172	AG	0.6	60	1.0
11176	AG	1.3	97	1.3
11177	CT	0.2	16	1.1
11197	CT	0.2	19	1.0
11204	CT	0.3	32	1.0
11215	CT	0.8	67	1.2
11233	CT	0.1	8	1.0
11242	CG	0.1	11	1.0
11251	AG	1.0	98	1.0
11253	CT	0.1	6	1.0
11255	CT	0.1	11	1.0
11257	CT	0.8	76	1.1
11269	CT	0.5	50	1.0
11287	CT	0.3	31	1.0
11288	CT	0.1	5	1.0
11293	AG	0.2	19	1.0
11296	CT	0.5	47	1.0
11299	CT	1.6	94	1.7
11302	CT	0.8	76	1.0
11314	AG	0.2	15	1.1
11332	CT	0.2	17	1.0
11339	CT	0.1	6	1.1
11348	CT	0.1	5	1.0
11353	CT	0.4	32	1.1
11362	AG	0.1	8	1.0
11365	CT	0.1	6	1.0
11368	CT	0.1	12	1.0
11377	AG	0.4	36	1.1
11383	CT	0.1	10	1.0
11404	AG	0.1	11	1.0
11410	CT	0.1	5	1.0
11413	AG	0.1	9	1.0
11431	CT	0.0	4	1.0
11437	CT	0.1	13	1.0
11440	AG	0.5	44	1.2
11447	CG	0.3	25	1.1
11467	AG	1.0	98	1.0
11470	AG	0.3	28	1.0
11476	CT	0.1	9	1.0
11482	CT	0.1	9	1.0
11485	CT	0.2	18	1.0
11533	CT	0.1	6	1.0
11536	CT	0.9	77	1.2
11547	CT	0.1	5	1.0
11549	CT	0.1	7	1.0
11554	CG	0.1	7	1.0
11569	CT	0.0	4	1.0
11590	AG	0.1	11	1.0
11611	AG	0.3	26	1.1
11620	AG	0.1	10	1.0
11623	CT	0.1	11	1.0
11641	AG	1.0	97	1.0
11647	CGT	0.8	73	1.1
11653	AG	0.5	50	1.0
11654	AG	0.7	72	1.0
11665	ACT	0.2	21	1.0
11674	CT	0.6	54	1.0
11696	AG	0.2	20	1.0
11701	CT	0.1	11	1.1
11719	AG	1.5	100	1.5
11722	CT	0.2	17	1.1
11732	CT	0.2	22	1.0
11800	ACG	0.1	11	1.1
11807	AG	0.1	8	1.0
11812	AG	1.0	76	1.3
11840	CT	0.1	6	1.0
11854	CT	0.2	24	1.0
11864	CT	0.2	17	1.0
11878	CT	0.1	6	1.0
11881	CT	0.1	11	1.0
11884	AG	0.1	7	1.3
11893	AG	0.1	5	1.0
11899	CT	1.1	96	1.1
11908	AG	0.0	4	1.0
11914	AG	4.6	100	4.6
11928	AG	0.1	12	1.0
11935	CT	0.4	33	1.1
11944	CT	1.5	100	1.5
11946	CT	0.0	4	1.0
11947	AG	0.5	54	1.0
11959	AG	0.2	18	1.1
11963	AG	0.3	28	1.0
11969	AG	0.5	45	1.2
11992	CT	0.1	10	1.0
12007	AG	2.2	100	2.2
12019	CT	0.5	53	1.0
12026	AG	0.9	69	1.3
12028	CT	0.1	5	1.0
12030	AG	0.1	9	1.0
12049	CT	0.7	72	1.0
12061	CT	0.1	5	1.2
12070	AG	0.5	47	1.0
12083	GT	0.1	7	1.0
12088	CT	0.1	9	1.0
12091	CT	0.2	21	1.0
12092	ACT	0.1	14	1.1
12106	CT	0.1	12	1.0
12121	CT	0.4	36	1.0
12127	AG	0.3	28	1.0
12135	ACT	0.1	11	1.0
12136	CT	0.0	3	1.0
12153	CT	0.1	15	1.0
12172	AG	1.4	83	1.7
12189	CT	0.4	35	1.1
12192	AG	0.3	28	1.1
12234	AG	0.4	32	1.2
12236	AG	1.3	97	1.3
12239	CT	0.5	54	1.0
12285	CT	0.1	11	1.0
12297	CT	0.1	8	1.0
12308	AG	1.0	98	1.0
12311	CT	0.3	33	1.0
12338	CT	0.2	16	1.0
12346	CT	0.1	10	1.1
12351	CT	0.3	32	1.0
12354	CT	0.1	11	1.0
12358	AG	1.2	79	1.5
12360	CT	0.1	7	1.0
12361	AG	0.6	55	1.0
12362	CT	0.1	6	1.1
12366	AG	0.1	11	1.1
12370	CT	0.1	7	1.0
12372	AG	1.6	99	1.6
12373	AG	0.0	2	1.0
12396	CT	0.2	21	1.1
12397	AG	0.1	11	1.2
12403	CT	0.0	4	1.0
12405	CT	0.7	71	1.0
12406	AG	1.2	87	1.4

E5.1 cont. Character scores for random 75-taxon minimal tree sets (page 7/9)

rCRS number	States	Ave. steps all data sets (n=114)	% of datasets pars. inform.	Ave. steps when inform.
12408	CT	0.2	18	1.0
12414	CT	0.4	39	1.1
12425	AG	0.1	8	1.1
12432	CT	0.4	32	1.1
12438	CT	0.1	12	1.0
12451	AG	0.0	3	1.0
12469	AG	0.0	4	1.0
12477	CT	0.6	48	1.2
12498	CT	0.1	7	1.0
12501	AG	1.3	81	1.7
12507	AG	0.3	29	1.1
12519	CT	1.1	100	1.1
12549	CT	0.1	15	1.0
12561	AG	0.2	21	1.0
12574	CT	0.1	6	1.0
12603	CT	0.0	4	1.0
12612	AGT	1.0	85	1.2
12613	AG	0.1	6	1.0
12616	CT	0.4	39	1.1
12618	AG	0.5	54	1.0
12630	AG	1.0	73	1.3
12633	ACT	0.9	67	1.3
12651	ACG	0.2	15	1.1
12654	AG	0.1	7	1.0
12669	CT	0.3	32	1.0
12672	AG	0.2	16	1.0
12684	AG	0.3	32	1.0
12693	AG	1.0	100	1.0
12696	CT	0.2	16	1.1
12705	ACT	1.7	100	1.7
12714	CT	0.1	15	1.0
12715	AG	0.1	5	1.0
12720	AG	1.1	97	1.1
12732	CT	0.5	47	1.1
12738	CGT	0.1	9	1.0
12753	AG	0.1	7	1.0
12768	AG	0.6	62	1.0
12771	AG	1.0	80	1.3
12793	CT	0.1	11	1.0
12795	AG	0.1	11	1.1
12810	AG	1.3	100	1.3
12811	CT	0.7	68	1.0
12816	CT	0.1	14	1.1
12822	AG	0.1	14	1.0
12842	CT	0.0	4	1.3
12850	AG	0.1	6	1.0
12858	CT	0.1	7	1.0
12864	CT	0.1	5	1.0
12870	CT	0.0	2	1.0
12876	CT	0.1	15	1.0
12879	CT	0.2	22	1.0
12880	CT	0.5	48	1.0
12882	CT	0.8	72	1.1
12909	AG	0.1	10	1.0
12930	AGT	0.8	75	1.1
12940	AG	0.4	35	1.0
12948	AG	0.9	88	1.0
12950	AG	0.1	6	1.0
12954	CT	0.3	26	1.0
12957	CT	0.2	18	1.2
12961	AG	0.2	21	1.0
12999	AG	0.1	8	1.0
13020	CT	0.8	64	1.3
13047	AG	0.0	4	1.0
13050	AG	0.1	14	1.0
13065	CT	0.0	4	1.0
13074	AG	0.1	10	1.0
13089	CT	0.1	5	1.0
13101	AC	0.1	12	1.1
13104	AG	0.7	57	1.2
13105	AG	2.1	100	2.1
13110	CT	0.1	10	1.0
13111	CT	0.1	11	1.0
13117	AG	0.1	10	1.0
13129	CT	0.1	7	1.0
13135	AG	0.5	41	1.1
13143	CT	0.2	21	1.1
13145	ACG	0.3	30	1.1
13149	AG	0.7	72	1.0
13152	AG	0.1	15	1.0
13183	AG	0.5	47	1.0
13188	CT	0.0	4	1.0
13194	AG	0.8	64	1.2
13197	CT	0.1	6	1.0
13215	CT	0.2	18	1.1
13221	AG	0.2	15	1.1
13236	AG	0.1	10	1.0
13254	CT	0.1	5	1.0
13260	CT	0.3	29	1.1
13263	AG	0.6	56	1.1
13269	AG	0.1	11	1.2
13276	AG	1.0	100	1.0
13278	AG	0.3	29	1.0
13281	CT	0.9	65	1.3
13287	CT	0.1	6	1.0
13293	CT	0.6	54	1.2
13326	CT	0.1	12	1.0
13350	AG	0.1	14	1.0
13359	AG	0.1	10	1.0
13368	AG	0.9	86	1.1
13383	CT	0.1	9	1.0
13395	AG	0.1	14	1.0
13404	CT	0.1	8	1.0
13422	AG	0.0	4	1.0
13434	AG	0.4	34	1.2
13437	CT	0.1	15	1.0
13452	CT	0.0	4	1.0
13461	CT	0.2	18	1.1
13469	AT	0.1	4	1.2
13477	AG	0.1	9	1.0
13485	AG	1.3	100	1.3
13500	CT	0.7	59	1.2
13506	CT	1.0	100	1.0
13512	AG	0.1	8	1.0
13542	AG	0.1	8	1.0
13557	AG	0.0	4	1.0
13563	AG	0.7	67	1.0
13590	AG	2.0	100	2.0
13602	CT	0.1	11	1.1
13606	AG	0.3	25	1.0
13614	AG	0.1	5	1.0
13617	CT	0.8	78	1.0
13626	CT	0.3	32	1.1
13629	AG	0.4	40	1.0
13635	CT	0.2	21	1.1
13637	AG	0.1	14	1.0
13641	CT	0.1	11	1.0
13650	CT	1.0	100	1.0
13651	AG	0.4	39	1.1
13656	CT	0.2	18	1.1
13680	CT	0.4	39	1.1
13681	AG	0.2	18	1.1
13708	AG	2.8	98	2.9
13710	AG	0.1	15	1.0
13716	AG	0.1	7	1.0
13722	AG	0.1	7	1.0
13734	CT	0.3	30	1.0
13740	CT	0.1	6	1.0
13743	CT	0.1	10	1.1
13748	AG	0.0	4	1.0
13749	CT	0.0	4	1.0
13754	CT	0.0	4	1.0
13758	CT	0.1	7	1.0
13759	AG	0.7	59	1.1
13768	ACT	0.2	17	1.0
13780	AG	0.4	38	1.0
13782	CT	0.1	11	1.0
13789	CT	1.0	100	1.0
13803	AG	1.0	100	1.0
13810	AG	0.3	26	1.0
13812	CT	0.1	9	1.0
13818	CT	0.0	4	1.0
13819	CT	0.5	48	1.0
13827	AG	0.6	47	1.2
13830	CT	0.1	13	1.0
13851	CT	0.0	4	1.0
13857	AG	0.1	13	1.0
13879	ACT	0.4	41	1.1
13880	AC	1.0	100	1.0
13886	CT	0.2	17	1.0
13887	AG	0.1	11	1.1
13889	AG	0.1	12	1.1
13890	CT	0.1	5	1.0
13899	CT	0.1	11	1.0
13914	AC	0.3	25	1.0
13924	CT	0.8	82	1.0
13928	ACGT	2.9	97	2.9
13934	CT	0.5	39	1.2
13942	AG	0.2	18	1.0
13958	CG	0.8	83	1.0
13965	CT	0.4	37	1.0
13966	AG	0.2	20	1.0
13980	AG	0.9	83	1.1
13981	CT	0.4	45	1.0
14000	AT	1.0	100	1.0
14002	AG	0.3	26	1.1
14007	AG	0.7	71	1.0
14013	AG	0.1	6	1.0
14016	AG	0.4	33	1.3
14020	CT	0.5	49	1.1
14022	AG	0.4	45	1.0
14025	CT	0.4	39	1.0
14034	CT	1.0	79	1.3
14040	AG	0.3	25	1.0
14053	AG	0.5	42	1.2
14059	AG	0.9	88	1.0
14063	CT	0.1	13	1.0
14070	AGT	0.1	15	1.0
14088	CT	1.0	93	1.1
14091	AG	0.2	17	1.0
14094	CT	0.2	19	1.0
14097	CT	0.1	12	1.0
14103	AC	0.1	8	1.0
14106	CT	0.3	26	1.0
14109	CT	0.3	27	1.1
14110	CT	0.0	3	1.0
14115	CT	0.1	9	1.0
14118	AG	0.6	52	1.2
14128	AG	0.4	38	1.1
14129	CT	0.0	4	1.0
14131	CT	0.3	30	1.0
14133	AG	0.4	39	1.0
14137	CT	0.1	8	1.0
14139	AG	0.2	16	1.0
14148	AG	1.2	98	1.3
14149	CT	0.1	10	1.0
14152	AG	0.3	32	1.1
14155	CT	0.1	10	1.1
14158	CT	0.1	11	1.1
14162	AG	0.1	9	1.0
14167	CT	0.9	85	1.0
14178	CT	1.2	100	1.2
14179	AG	0.3	29	1.1
14180	CT	0.8	54	1.4
14182	CT	1.2	79	1.5
14185	AT	0.1	11	1.0
14194	CT	0.1	10	1.0
14198	AG	0.1	7	1.0
14200	CT	0.7	65	1.0
14203	AG	1.0	100	1.0
14209	AG	0.1	8	1.0
14212	CT	0.5	46	1.0
14215	CT	0.1	9	1.0
14233	AG	0.8	75	1.1
14281	CGT	0.3	29	1.1
14284	CT	0.2	16	1.0
14287	CT	0.4	38	1.0
14290	CT	0.1	11	1.0

E5.1 cont. Character scores for random 75-taxon minimal tree sets (page 8/9)

rCRS number	States	Ave. steps all data sets (n=114)	% of datasets pars. inform.	Ave. steps when inform.
14299	CT	0.3	31	1.0
14302	CT	0.0	4	1.0
14305	AG	0.5	46	1.1
14308	CT	1.3	98	1.3
14311	CT	0.3	28	1.0
14318	CT	0.5	51	1.0
14323	AG	0.2	24	1.0
14325	CT	0.2	15	1.1
14337	CT	0.1	8	1.0
14338	CT	0.3	32	1.1
14340	CT	0.0	4	1.0
14364	AG	0.8	73	1.1
14365	CT	0.2	20	1.0
14370	AG	0.1	8	1.0
14371	CT	0.6	52	1.1
14374	CT	0.8	63	1.3
14384	ACG	0.2	18	1.0
14393	AG	0.3	25	1.0
14407	CT	0.1	15	1.0
14417	AG	0.1	10	1.1
14470	ACT	1.2	77	1.5
14476	AG	0.4	43	1.0
14484	CT	0.4	42	1.0
14488	CT	0.1	6	1.1
14502	CT	0.4	36	1.1
14527	AG	0.1	7	1.0
14533	CT	0.0	4	1.0
14544	AG	0.8	65	1.2
14548	AG	0.1	6	1.0
14550	CT	0.1	4	1.2
14552	AG	0.0	4	1.0
14560	ACG	1.3	100	1.3
14566	AG	1.2	100	1.2
14569	AG	1.1	92	1.2
14577	CT	0.1	10	1.1
14581	CT	0.1	7	1.0
14582	AG	0.2	24	1.0
14587	AG	0.1	5	1.0
14599	AG	0.5	51	1.0
14605	AG	0.6	61	1.0
14620	CT	0.4	36	1.1
14629	CT	0.1	7	1.0
14668	CT	1.2	100	1.2
14674	CT	0.1	8	1.0
14687	AG	0.3	29	1.0
14692	AG	0.2	16	1.1
14693	AG	0.3	28	1.0
14696	AG	0.1	12	1.1
14750	AG	0.2	17	1.1
14755	AGT	0.3	29	1.1
14757	CT	0.0	4	1.0
14766	CT	1.6	100	1.6
14769	AG	1.1	100	1.1
14770	CT	0.1	7	1.0
14783	CT	1.0	100	1.0
14790	AG	0.1	5	1.0
14793	AG	0.6	48	1.2
14794	CT	0.4	45	1.0
14798	CT	1.8	99	1.8
14812	CT	0.5	48	1.1
14831	AG	0.5	42	1.1
14861	AG	0.1	12	1.1
14862	CT	0.1	9	1.0
14867	CT	0.1	11	1.0
14869	ACG	0.1	11	1.1
14870	AG	0.1	10	1.0
14872	CT	0.2	23	1.0
14890	AG	0.2	18	1.0
14893	AG	0.5	47	1.0
14905	AG	1.1	86	1.3
14911	CT	1.0	100	1.0
14914	AG	0.1	10	1.0
14927	AG	0.6	50	1.1
14935	CT	0.1	5	1.0
14944	CT	0.6	55	1.0
14959	AGT	0.1	6	1.0
14968	CT	0.0	4	1.0
14971	CT	0.0	4	1.0
14978	AG	0.0	4	1.0
14979	CT	0.7	66	1.0
14990	CT	0.2	17	1.1
14996	AG	0.1	6	1.0
15016	CT	0.6	61	1.0
15022	CT	0.0	4	1.0
15040	CT	0.3	28	1.1
15043	AG	1.7	100	1.7
15047	AG	0.2	17	1.0
15049	CT	0.1	12	1.0
15055	CT	0.1	7	1.0
15061	AG	0.1	10	1.2
15067	CT	0.8	56	1.3
15071	CT	0.1	15	1.0
15090	CT	0.1	6	1.0
15099	CT	0.4	33	1.2
15100	CT	0.1	5	1.0
15106	AG	0.3	26	1.1
15110	AG	1.2	98	1.2
15115	CT	1.1	100	1.1
15136	CT	1.0	97	1.0
15148	AG	0.4	33	1.1
15172	AG	0.4	32	1.2
15184	CT	0.2	16	1.0
15191	CT	0.1	11	1.2
15204	CT	0.4	35	1.1
15211	CT	0.5	54	1.0
15214	CT	0.1	5	1.0
15217	AG	1.2	98	1.2
15218	ACG	0.5	46	1.1
15221	AG	0.1	11	1.0
15223	CT	0.4	44	1.0
15226	AG	0.7	66	1.0
15229	CT	0.7	66	1.0
15235	AG	0.2	22	1.0
15236	AG	1.3	80	1.6
15244	AG	1.8	100	1.8
15257	AG	0.1	14	1.1
15258	AG	0.1	5	1.2
15259	CT	0.1	6	1.0
15261	AG	0.2	19	1.0
15262	CT	0.1	7	1.0
15292	CT	0.1	10	1.0
15300	CT	0.1	11	1.0
15301	AG	2.5	100	2.5
15307	CT	0.1	10	1.0
15311	AG	0.2	20	1.0
15314	AG	0.2	21	1.1
15317	AG	0.1	10	1.1
15323	AG	0.6	59	1.1
15325	ACG	0.1	5	1.0
15326	AG	0.7	60	1.2
15331	AC	0.1	13	1.0
15340	AG	0.0	3	1.0
15341	CT	0.1	13	1.0
15346	AG	0.6	57	1.0
15355	AG	0.3	27	1.2
15385	CT	0.2	19	1.0
15391	CT	0.7	68	1.1
15402	CT	0.1	6	1.0
15412	CT	0.1	8	1.0
15421	AG	0.5	54	1.0
15422	AG	0.1	7	1.0
15431	AG	1.4	98	1.4
15440	CT	0.7	58	1.1
15448	CT	0.1	6	1.0
15451	CT	0.4	38	1.0
15452	AC	1.0	98	1.0
15454	CT	0.2	20	1.1
15460	CT	0.1	5	1.0
15466	AG	0.3	32	1.1
15470	CT	0.3	27	1.1
15475	AG	0.1	11	1.0
15479	CT	0.1	15	1.0
15487	AGT	0.9	84	1.1
15496	AG	0.1	7	1.0
15497	AG	1.0	68	1.4
15508	CT	0.4	45	1.0
15511	CT	0.2	20	1.1
15514	CT	0.1	13	1.1
15518	CT	0.4	43	1.0
15519	CT	0.2	18	1.1
15524	AG	0.5	47	1.0
15530	CT	0.3	27	1.1
15535	CT	0.6	59	1.0
15553	AG	0.1	7	1.0
15562	AG	0.5	42	1.1
15565	CT	0.0	4	1.0
15589	AC	0.1	10	1.0
15601	CT	0.1	7	1.0
15607	AG	1.4	93	1.5
15613	AG	0.1	10	1.0
15622	CT	0.4	40	1.0
15626	CT	0.3	26	1.0
15629	CT	1.1	100	1.1
15632	ACT	0.2	18	1.0
15661	CT	0.1	6	1.0
15662	AG	0.4	45	1.0
15663	CT	0.7	61	1.1
15664	CT	0.1	8	1.1
15670	CT	0.6	47	1.2
15674	CT	0.0	4	1.0
15682	AG	0.1	5	1.0
15693	CT	0.2	16	1.0
15697	CT	0.5	41	1.1
15721	CT	0.2	16	1.0
15724	AG	0.3	32	1.1
15731	AG	0.1	5	1.0
15734	AG	0.4	39	1.1
15746	AG	0.6	57	1.1
15748	CT	0.1	9	1.0
15752	AG	0.0	4	1.0
15754	ACT	0.1	13	1.0
15758	AG	0.9	59	1.5
15766	AG	0.2	24	1.0
15769	ACG	0.1	11	1.1
15773	AG	0.1	10	1.0
15777	ACG	0.3	30	1.0
15784	CT	2.1	100	2.1
15790	CT	0.0	4	1.0
15793	CT	0.1	9	1.0
15799	AG	0.0	4	1.0
15805	AG	0.2	17	1.0
15806	AG	0.1	9	1.0
15812	AG	0.1	15	1.0
15817	AG	0.1	5	1.0
15824	AG	0.3	25	1.0
15833	CT	0.4	37	1.0
15849	CT	0.9	83	1.0
15850	CGT	0.3	32	1.0
15851	AG	0.4	45	1.0
15852	CT	0.3	31	1.1
15860	AG	0.4	43	1.0
15865	AG	0.1	10	1.0
15874	AG	0.6	57	1.0
15883	AG	0.1	10	1.0
15884	ACG	0.9	66	1.3
15885	CT	0.1	11	1.0
15886	CT	0.1	8	1.0
15889	CT	0.3	31	1.0
15900	CT	0.1	14	1.1
15904	CT	0.8	78	1.0
15905	CT	0.8	76	1.0
15907	AG	0.3	28	1.0
15908	CT	0.0	4	1.0
15924	AG	2.5	96	2.6
15927	AG	0.5	51	1.1
15928	AG	0.9	86	1.1
15930	AG	0.6	45	1.3
15932	CT	0.1	10	1.2

E5.1 cont. Character scores for random 75-taxon minimal tree sets (page 9/9)

rCRS number	States	Ave. steps all data sets (n=114)	% of datasets pars. inform.	Ave. steps when inform.
15937	AT	0.1	9	1.0
15938	CT	0.2	21	1.0
15940	CT	0.1	11	1.1
15941	CT	0.5	41	1.2
15942	CT	0.3	26	1.1
15946	CT	0.1	9	1.0
15951	AG	0.5	52	1.1
15954	ACG	0.6	49	1.2
15968	CT	0.1	12	1.0
15978	CT	0.8	76	1.0
16000	AG	0.0	4	1.0
16017	CT	0.3	29	1.1
16048	AG	0.1	12	1.1
16051	AG	1.6	81	2.0
16066	AG	0.2	16	1.2
16067	CT	0.1	5	1.0
16069	CT	0.9	85	1.0
16071	CT	0.3	30	1.0
16075	CT	0.2	21	1.1
16077	AT	0.1	7	1.0
16086	CT	2.6	96	2.7
16092	CT	1.0	67	1.5
16093	CT	3.8	99	3.9
16094	CT	0.1	5	1.0
16095	CT	0.1	10	1.0
16102	CT	0.1	10	1.0
16108	CT	0.2	17	1.0
16111	ACT	1.4	79	1.7
16114	ACGT	2.2	96	2.3
16124	CT	0.4	37	1.1
16126	CT	2.6	100	2.6
16129	ACG	9.7	100	9.7
16136	CT	0.6	51	1.1
16140	CT	0.9	70	1.4
16144	ACT	0.6	46	1.2
16145	AG	2.3	96	2.4
16146	AG	0.1	12	1.1
16147	CT	0.1	14	1.1
16148	CT	1.5	99	1.5
16150	CT	0.1	12	1.1
16153	AG	0.7	51	1.3
16154	CT	0.2	18	1.1
16160	AG	0.0	4	1.0
16162	AG	0.8	65	1.3
16163	AG	0.6	54	1.2
16164	AGT	0.4	35	1.1
16167	CT	0.2	15	1.1
16168	CT	1.2	89	1.3
16169	ACT	0.5	43	1.2
16170	ACG	0.1	14	1.1
16171	AG	0.0	4	1.0
16172	CT	4.6	100	4.6
16173	CT	0.3	27	1.1
16174	CT	0.3	30	1.1
16175	AG	0.1	5	1.0
16176	ACGT	0.6	47	1.3
16178	CT	0.1	14	1.0
16180	AG	0.1	13	1.0
16182	ACGT	3.2	100	3.2
16183	ACG	6.6	100	6.6
16184	ACT	1.0	65	1.5
16185	CT	0.9	65	1.4
16186	CT	0.4	41	1.1
16187	CT	3.6	100	3.6
16188	ACGT	1.8	97	1.9
16189	CT	12.7	100	12.7
16190	CT	0.1	7	1.0
16194	ACG	0.6	41	1.4
16195	CT	0.5	38	1.2
16203	AG	0.1	12	1.0
16206	AC	0.1	9	1.0
16207	AG	0.1	10	1.0
16209	CT	2.3	91	2.5
16212	AG	0.5	46	1.2
16213	AG	2.0	97	2.0
16214	CT	1.6	93	1.8
16215	ACGT	1.2	75	1.6
16216	ACG	0.1	11	1.0
16217	CT	2.0	97	2.1
16218	CT	0.3	28	1.1
16219	AG	0.4	39	1.0
16222	CT	0.4	35	1.1
16223	CT	4.7	100	4.7
16224	CT	1.3	90	1.5
16227	AG	0.7	59	1.1
16229	CT	0.3	28	1.0
16230	AG	1.1	100	1.1
16231	CT	0.4	32	1.2
16232	ACT	0.5	47	1.1
16233	AG	0.1	5	1.0
16234	CT	2.9	99	2.9
16235	AG	0.4	32	1.2
16239	CGT	0.8	61	1.3
16240	ACG	0.1	8	1.0
16241	AGT	0.6	51	1.2
16242	ACT	0.7	61	1.2
16243	CT	0.9	64	1.4
16244	AG	0.2	18	1.1
16245	CT	0.7	61	1.2
16247	AG	0.4	33	1.1
16248	CT	0.1	9	1.0
16249	CT	1.7	93	1.8
16250	CT	0.1	10	1.0
16254	ACG	0.1	6	1.0
16255	AG	0.1	8	1.0
16256	CT	0.9	64	1.5
16257	ACT	0.7	65	1.1
16258	ACGT	0.5	41	1.1
16259	CT	0.1	10	1.2
16260	ACT	1.0	68	1.4
16261	CT	2.2	96	2.3
16263	CT	0.2	24	1.0
16264	CT	1.7	100	1.7
16265	ACGT	1.3	88	1.5
16266	ACGT	1.2	75	1.6
16268	CT	0.1	14	1.0
16269	AG	0.3	26	1.0
16270	CT	2.3	100	2.3
16271	CT	0.5	40	1.2
16272	AG	0.0	3	1.0
16274	AG	2.2	96	2.3
16278	CT	4.7	100	4.7
16284	AG	0.9	69	1.3
16286	ACGT	2.4	96	2.5
16287	CT	0.8	64	1.2
16288	CT	0.4	36	1.2
16289	AG	0.1	5	1.0
16290	CT	1.3	93	1.4
16291	CGT	2.6	96	2.7
16292	CGT	1.7	84	2.1
16293	ACG	4.0	100	4.0
16294	CT	4.0	100	4.0
16295	CT	0.8	58	1.4
16296	CT	0.7	69	1.1
16297	CT	0.8	72	1.1
16298	CT	2.9	100	2.9
16299	AG	0.2	23	1.0
16300	AG	0.4	35	1.2
16301	CT	0.2	19	1.0
16302	AG	0.2	20	1.0
16303	AG	0.1	11	1.0
16304	CT	2.6	98	2.7
16309	AG	2.6	100	2.6
16311	CT	8.1	100	8.1
16316	AG	0.4	35	1.1
16317	AGT	0.2	15	1.1
16318	ACGT	0.6	43	1.4
16319	AG	3.1	100	3.1
16320	CT	1.7	97	1.8
16324	CT	0.9	75	1.2
16325	CT	1.1	71	1.5
16326	AG	0.0	4	1.0
16327	CT	1.0	69	1.5
16335	AG	0.2	17	1.1
16337	CT	0.1	5	1.0
16342	CT	0.1	7	1.0
16343	ACG	0.3	29	1.2
16344	CT	0.5	50	1.1
16352	CT	0.4	34	1.2
16353	CT	0.1	9	1.0
16354	CT	0.5	43	1.1
16355	ACT	1.6	90	1.7
16356	CT	0.8	59	1.4
16357	CT	0.6	50	1.2
16359	CT	0.1	13	1.1
16360	ACT	1.4	100	1.4
16362	ACT	7.0	100	7.0
16366	CT	0.1	9	1.0
16368	CT	1.3	82	1.6
16381	CT	0.2	18	1.0
16390	AG	3.4	100	3.4
16391	AG	0.4	36	1.0
16399	AG	2.0	95	2.1
16400	CT	0.4	32	1.2
16428	AG	0.1	10	1.1
16438	AG	0.1	7	1.0
16463	AG	0.4	37	1.0
16465	CT	0.1	8	1.0
16468	CT	0.2	17	1.1
16470	AG	0.1	10	1.0
16471	AG	0.1	10	1.0
16482	AG	0.1	9	1.0
16497	AG	1.0	64	1.5
16519	CT	18.9	100	18.9
16523	AG	0.0	4	1.0
16524	AG	0.4	35	1.2
16526	AG	0.4	35	1.2
16527	CT	1.0	77	1.2

E6.1 Haplotypes from HVR-I nt16065-nt16373 data set found in Oceania (page 1/6)

This table contains details of the 221 haplotypes from the HVR-I nt16065-nt16373 data set which included sequences from Oceania. It is ordered firstly by the tentatively assigned haplogroups, and then by haplotype name. The 'Regions' column is shaded for haplotypes which are found in more than one region and the number of sequences from each region is shown in parentheses. The differences to the rCRS for each haplotype were obtained using Sequencher™ (Gene Codes Corporation).

Haplotype	Differences to rCRS	n	Haplogroup	Regions	Entire mt genomes
AB119302Balo	16182CtvA 16183CtvA 16189C 16217C 16247G 16261T 16293G	3	B4a	Other Near Oceania	
AB119304Balo	16182CtvA 16183CtvA 16189C 16217C	7	B4a	East Asia (2), Americas (4), Other Near Oceania (1)	
AY289083PNG	16093C 16182CtvA 16183CtvA 16189C 16217C 16261T	5	B4a	East Asia (4), New Guinea (1)	AY289083B4a1a1
DQ372873Trob	16129A 16182CtvA 16183CtvA 16189C 16217C 16261T	19	B4a	East Asia (16), Taiwan (2), New Guinea (1)	DQ372873B4a1a1
EF077363NZFiji	16182CtvA 16183CtvA 16189C 16217C 16261T 16311C	10	B4a	Taiwan (8), Remote Oceania (2)	
EF077392NZTong	16182CtvA 16183CtvA 16189C 16217C 16261T	53	B4a	East Asia (5), Taiwan (27), Indonesia (1), Malaysia (1), Philippines (3), New Guinea (3), Remote Oceania (13)	AJ842745, AY289076, DQ372871, DQ372874, DQ372875, DQ372877 allB4a1a1
AB119299Balo	16182CtvA 16183CtvA 16189C 16217C 16246G 16247G	1	B4a1a1PM	Other Near Oceania	
AB119303Balo	16182CtvA 16183CtvA 16189C 16217C 16247G	6	B4a1a1PM	Other Near Oceania	
AB119337Balo	16182CtvA 16183CtvA 16189C 16217C 16242AtvC 16247G 16261T	2	B4a1a1PM	Other Near Oceania	
AB119345Haap	16182CtvA 16183CtvA 16189C 16217C 16218T 16247G 16261T	1	B4a1a1PM	Remote Oceania	
AB119352Haap	16182CtvA 16183CtvA 16189C 16217C 16247G 16248T 16261T	1	B4a1a1PM	Remote Oceania	
AB119366Tong	16182CtvA 16183CtvA 16189C 16217C 16247G 16261T 16342C	1	B4a1a1PM	Remote Oceania	
AB119373Tong	16182CtvA 16183CtvA 16189C 16217C 16247G 16261T 16362C	1	B4a1a1PM	Remote Oceania	
AB119377Tong	16182CtvA 16183CtvA 16189C 16217C 16247G 16261T 16278T	1	B4a1a1PM	Remote Oceania	
AB119381Tong	16126C 16182CtvA 16183CtvA 16189C 16217C 16247G 16261T	1	B4a1a1PM	Remote Oceania	
AB119385Tong	16142T 16182CtvA 16183CtvA 16189C 16217C 16247G 16261T	1	B4a1a1PM	Remote Oceania	
AF347007Samo	16182CtvA 16183CtvA 16189C 16217C 16247G 16261T 16324C	1	B4a1a1PM	Remote Oceania	AF347007B4a1a1PM
AY289069Cook	16182CtvA 16183CtvA 16189C 16217C 16247G 16261T 16317G	1	B4a1a1PM	Remote Oceania	
AY289094Samo	16182CtvA 16183CtvA 16189C 16217C 16247G 16261T 16352C	1	B4a1a1PM	Remote Oceania	AY289094B4a1a1PM
AY289102Tong	16182CtvA 16183CtvA 16189C 16217C 16247G 16261T 16360T	1	B4a1a1PM	Remote Oceania	AY289102B4a1a1PM
AY604118NZSamo	16111T 16182CtvA 16183CtvA 16189C 16217C 16247G 16261T	1	B4a1a1PM	Remote Oceania	
AY604140NZMaor	16182CtvA 16183CtvA 16189C 16217C 16242T 16247G 16261T	1	B4a1a1PM	Remote Oceania	
AY604153NZNiue	16181G 16182CtvA 16183CtvA 16189C 16217C 16247G 16261T	1	B4a1a1PM	Remote Oceania	
DQ309848Kark	16183CtvA 16189C 16217C 16247G 16261T	1	B4a1a1PM	New Guinea	
DQ309849Kark	16182CtvA 16183CtvA 16189C 16247G 16261T	1	B4a1a1PM	New Guinea	
DQ309850Kark	16181G 16182CtvA 16183CtvA 16189C 16247G 16261T 16357C	1	B4a1a1PM	New Guinea	
DQ309851Kark	16126C 16181G 16182CtvA 16183CtvA 16189C 16247G 16261T 16357C	1	B4a1a1PM	New Guinea	
DQ309852Kark	16126C 16182CtvA 16183CtvA 16189C 16247G 16261T	1	B4a1a1PM	New Guinea	
EF077361NZMaor	16086C 16182CtvA 16183CtvA 16189C 16217C 16247G 16261T	1	B4a1a1PM	Remote Oceania	
EF077364NZSamo	16182CtvA 16183CtvA 16189C 16217C 16247G 16261T 16263C	1	B4a1a1PM	Remote Oceania	
EF077365NZMaor	16092C 16182CtvA 16183CtvA 16189C 16217C 16247G 16261T	3	B4a1a1PM	Remote Oceania	
EF077370NZTong	16182CtvA 16183CtvA 16189C 16217C 16247G 16261T 16311C	7	B4a1a1PM	Other Near Oceania (3), Remote Oceania (4)	AY963574B4a1a1PM
EF077372NZSamo	16182CtvA 16183CtvA 16189C 16217C 16247G 16249C 16261T	1	B4a1a1PM	Remote Oceania	
EF077381NZCook	16182CtvA 16183CtvA 16189C 16217C 16247G 16261T 16291T	2	B4a1a1PM	Remote Oceania	
EF077383NZSamo	16181CtvA 16182CtvA 16183CtvA 16189C 16217C 16247G 16261T	1	B4a1a1PM	Remote Oceania	
EF077398NZTong	16163G 16182CtvA 16183CtvA 16189C 16217C 16247G 16261T	2	B4a1a1PM	Remote Oceania	
EF077400NZSamo	16093C 16182CtvA 16183CtvA 16189C 16217C 16247G 16261T	1	B4a1a1PM	Remote Oceania	
EF077402NZCook	16182CtvA 16183CtvA 16189C 16217C 16247G 16261T	86	B4a1a1PM	New Guinea (2), Other Near Oceania (22), Remote Oceania (62)	AY289068, AY289077, AY289078, AY289080, AY289093, DQ372878, DQ372881, DQ372886 All B4a1a1PM

E6.1 cont. Haplotypes from HVR-I nt16065-nt16373 data set found in Oceania (page 2/6)

Haplotype	Differences to rCRS	n	Haplogroup	Regions	Entire mt genomes
DQ137411Bour	16077TtvA 16136C 16172C 16183CtvA 16189C 16223T 16311C 16320T	2	M27a	Other Near Oceania	DQ137410, DQ137411M/M27a
DQ137404Bism	16145A 16209C 16299G	3	M27b	Other Near Oceania	DQ137402, DQ137403, DQ137404M/M27b
DQ137406Bism	16223T 16301T 16304C	2	M27c	Other Near Oceania	DQ137405, DQ137406M/M27c
U47175Vanu1	16148T 16223T 16362C	1	M28	Remote Oceania	
U47176Mars2	16148T 16223T 16295T 16362C	2	M28	Remote Oceania	
DQ137400Bism	16086C 16129A 16148T 16223T 16320T 16362C	1	M28a	Other Near Oceania	DQ137400M/M28a
DQ137401Bism	16086C 16129A 16148T 16223T 16362C	1	M28a	Other Near Oceania	DQ137401M/M28a
DQ372879Vanu	16086C 16129A 16148T 16223T 16362C 16366T	3	M28a	Remote Oceania	DQ372879M/M28a
DQ372883Vanu	16086C 16129A 16148T 16223T 16311C 16362C	1	M28a	Remote Oceania	DQ372883M/M28a
U47171Vanu2	16129A 16148T 16223T 16311C 16343G	2	M28a	Remote Oceania	
U47172Vanu7	16129A 16148T 16223T 16343G	7	M28a	Remote Oceania	
U47173Vanu1	16129A 16148T 16223T 16355T 16362C	1	M28a	Remote Oceania	
U47174Vanu1	16129A 16148T 16223T 16355T 16362C 16365T	1	M28a	Remote Oceania	
DQ137398Bism	16093C 16148T 16223T 16318TtvA 16362C	1	M28b	Other Near Oceania	DQ137398M/M28b
DQ137399Bism	16148T 16223T 16291T 16318TtvA 16362C	1	M28b	Other Near Oceania	DQ137399M/M28b
DQ137409Bism	16182CtvA 16183CtvA 16189C 16223T 16311C	3	M29	Other Near Oceania	DQ137407, DQ137408, DQ137409M/M29
U47217Mars1	16223T 16295T	11	M7	East Asia (4), Borneo (4), Taiwan (1), Remote Oceania (2)	AY255158M/M7
DQ372876Mars	16223T 16295T 16362C	79	M7c	Borneo (4), Taiwan (19), Indonesia (38), Malaysia (1), Philippines (16), Remote Oceania (1)	DQ372876, AF382012M/M7c
U47189Cook1	16093C 16223T 16298C 16327T	5	M8	East Asia (4), Remote Oceania (1)	AY519490M/M8
AB119326Balo	16176T 16266T 16357C	9	P1	New Guinea (7), Other Near Oceania (2)	
AB119440Gidr	16176T 16189C 16266T 16270T 16357C	1	P1	New Guinea	
AF347004PNG	16111AtvC 16169T 16266T 16294T 16357C	1	P1	New Guinea	AF347007N/R/P1
AF347005PNG	16176T 16183CtvA 16189C 16223T 16235G 16257T 16266T 16270T 16354T 16357C	1	P1	New Guinea	AF347005N/R/P1
AJ635016Iria	16110A 16176T 16266T 16357C	1	P1	New Guinea	
AJ635041Iria	16176T 16213A 16266T 16270T 16357C	1	P1	New Guinea	
AJ635067Iria	16093C 16176T 16225T 16266T 16270T 16318G 16357C	2	P1	New Guinea	
AJ635119Iria	16266T 16270T 16357C	1	P1	New Guinea	
AJ635121Iria	16093C 16176T 16209C 16266T 16357C	1	P1	New Guinea	
AJ635133Iria	16110A 16176T 16223T 16262T 16266T 16270T 16311C 16357C	1	P1	New Guinea	
AJ635142Iria	16176T 16223T 16262T 16266T 16270T 16311C 16357C	5	P1	New Guinea	
AJ635154Iria	16176T 16221T 16266T 16270T 16311C 16357C	6	P1	New Guinea	
AY289092PNG	16176T 16266T 16291T 16292T 16335G 16357C	1	P1	New Guinea	AY289092N/R/P1
U25385PNG	16176T 16257T 16266T 16270T 16354T 16357C	2	P1	New Guinea	AY289087N/R/P1
U25386PNG	16176T 16235G 16257T 16264T 16266T 16270T 16304C 16354T 16357C	1	P1	New Guinea	
U25387PNG	16176T 16235G 16266T 16357C	2	P1	New Guinea	AF347002N/R/P1
U25388PNG	16176T 16266T 16287T 16294T 16357C	2	P1	New Guinea	AY289086N/R/P1
U47220Vanu1	16266T 16357C	1	P1	Remote Oceania	
U47253Vanu2	16176T 16209C 16266T 16357C	7	P1	New Guinea (5), Remote Oceania (2)	
U47254Vanu3	16176T 16266T 16270T 16357C	4	P1	New Guinea (1), Remote Oceania (3)	
U47260Vanu1	16266T 16291T 16311C 16357C	1	P1	Remote Oceania	
U47265Vanu1	16176T 16209C 16266T 16274A 16357C	1	P1	Remote Oceania	
U47269Vanu1	16176T 16250T 16266T 16291T 16311C 16357C	1	P1	Remote Oceania	
AY289088PNG	16188T 16278T 16357C	2	P2	New Guinea	AY289088, AY289084N/R/P2

E6.1 cont. Haplotypes from HVR-I nt16065-nt16373 data set found in Oceania (page 3/6)

Haplotype	Differences to rCRS	n	Haplogroup	Regions	Entire mt genomes
U25389PNG	16145A	4	P3	New Guinea	AY289091N/R/P3
AJ635116Iria	16093C 16129A 16148T 16163G 16214T 16223T 16231C 16241G 16311C	1	Q	New Guinea	
AJ635137Iria	16129A 16148T 16163G 16214T 16223T 16241G 16311C	2	Q	New Guinea	
DQ309859Kark	16129A 16223T 16241G	1	Q	New Guinea	
U25382PNG	16129A 16145A 16150T 16223T 16241G 16311C	1	Q	New Guinea	
U25383PNG	16129A 16223T 16241G 16311C	3	Q	Indonesia (1), New Guinea (2)	
U47243Vanu1	16223T 16241G 16319A	1	Q	Remote Oceania	
U47271Vanu1	16223T 16234T 16241G 16269G 16292T 16297C 16311C 16362C	1	Q	Remote Oceania	
AB119306Balo	16129A 16144C 16148T 16223T 16261T 16265CtvA 16311C 16343G	2	Q1	New Guinea (1), Other Near Oceania (1)	
AB119308Balo	16129A 16144C 16148T 16174T 16223T 16241G	2	Q1	Other Near Oceania	
AB119336Balo	16129A 16144C 16148T 16172C 16174T 16223T 16241G 16259AtvC 16265CtvA 16300G 16311C 16343G	3	Q1	Other Near Oceania	
AJ634998Iria	16129A 16144C 16148T 16223T 16261T 16265CtvA 16291T 16298C 16311C 16343G	7	Q1	New Guinea	
AJ635000Iria	16241G 16243C 16265CtvA 16311C 16343G	1	Q1	New Guinea	
AJ635006Iria	16092C 16129A 16144C 16148T 16223T 16241G 16265CtvA 16273A 16274A 16304C 16311C 16343G	1	Q1	New Guinea	
AJ635007Iria	16129A 16144C 16148T 16218T 16223T 16241G 16243C 16265CtvA 16311C 16343G	1	Q1	New Guinea	
AJ635008Iria	16129A 16144C 16148T 16223T 16234T 16241G 16265CtvA 16311C 16343G	1	Q1	New Guinea	
AJ635011Iria	16129A 16144C 16148T 16223T 16263C 16265CtvA 16311C 16343G	1	Q1	New Guinea	
AJ635012Iria	16129A 16144C 16148T 16223T 16261T 16265CtvA 16291T 16298C 16299G 16311C 16343G 16373A	1	Q1	New Guinea	
AJ635013Iria	16129A 16144C 16148T 16223T 16261T 16265CtvA 16291T 16298C 16299G 16311C 16343G	1	Q1	New Guinea	
AJ635019Iria	16129A 16144C 16148T 16172C 16223T 16224 16241G 16265CtvA 16311C 16343G	1	Q1	New Guinea	
AJ635043Iria	16110A 16129A 16148T 16223T 16241G 16265CtvA 16292T 16311C 16343G	2	Q1	New Guinea	
AJ635045Iria	16129A 16144C 16148T 16172C 16209C 16223T 16241G 16265CtvA 16311C 16343G	4	Q1	New Guinea	
AJ635048Iria	16129A 16144C 16148T 16222T 16241G 16265CtvA 16311C 16343G	6	Q1	New Guinea	
AJ635051Iria	16129A 16144C 16148T 16172C 16223T 16241G 16265CtvA 16290T 16311C 16343G	2	Q1	New Guinea	
AJ635055Iria	16129A 16148T 16222T 16241G 16265CtvA 16311C 16343G	5	Q1	New Guinea	
AJ635057Iria	16129A 16144C 16148T 16223T 16261T 16265CtvA 16291T 16298C 16311C 16343G 16373A	1	Q1	New Guinea	
AJ635060Iria	16129A 16144C 16148T 16223T 16227G 16241G 16265CtvA 16311C 16343G	2	Q1	New Guinea	
AJ635073Iria	16129A 16144C 16148T 16213A 16223T 16241G 16265CtvA 16311C 16343G	3	Q1	New Guinea	
AJ635098Iria	16129A 16144C 16148T 16223T 16241G 16265CtvA 16270T 16311C 16343G	14	Q1	Indonesia (2), New Guinea (12)	
AJ635099Iria	16129A 16144C 16148T 16223T 16241G 16265CtvA 16270T 16311C 16343G	1	Q1	New Guinea	
AJ635105Iria	16129A 16144C 16148T 16162G 16174T 16223T 16224C 16241G 16265CtvA 16311C 16343G	1	Q1	New Guinea	
AJ635106Iria	16067T 16129A 16144C 16148T 16223T 16241G 16265CtvA 16311C 16343G	1	Q1	New Guinea	
AJ635109Iria	16223T 16241G 16265CtvA 16311C 16343G	1	Q1	New Guinea	
AJ635112Iria	16129A 16144C 16148T 16223T 16224C 16241G 16265CtvA 16311C 16343G	1	Q1	New Guinea	
AJ635114Iria	16129A 16144C 16148T 16222T 16224C 16241G 16265CtvA 16304C 16311C 16343G	1	Q1	New Guinea	
AJ635118Iria	16092C 16129A 16144C 16148T 16223T 16265CtvA 16305G 16311C 16343G	1	Q1	New Guinea	
AJ635125Iria	16129A 16144C 16148T 16172C 16223T 16265CtvA 16311C 16343G	1	Q1	New Guinea	
AJ635143Iria	16092C 16129A 16144C 16148T 16223T 16241G 16265CtvA 16274A 16311C 16343G	2	Q1	New Guinea	
AJ635153Iria	16129A 16144C 16148T 16223T 16241G 16243C 16265CtvA 16311C 16343G	3	Q1	New Guinea	

E6.1 cont. Haplotypes from HVR-I nt16065-nt16373 data set found in Oceania (page 4/6)

Haplotype	Differences to rCRS	n	Haplogroup	Regions	Entire mt genomes
AJ635156Iria	16093C 16129A 16144C 16148T 16223T 16241G 16265CtvA 16311C 16343G	20	Q1	New Guinea	
AJ635158Iria	16129A 16144C 16148T 16172C 16223T 16241G 16265CtvA 16270T 16311C 16343G	14	Q1	Borneo (1), Indonesia (4), New Guinea (9)	
AJ635159Iria	16144C 16148T 16223T 16241G 16265CtvA 16270T 16311C	2	Q1	New Guinea	
AJ635160Iria	16129A 16148T 16241G 16265CtvA 16311C 16343G	1	Q1	New Guinea	
AY289075Bour	16129A 16144C 16148T 16241G 16265CtvA 16311C 16343G	8	Q1	Indonesia (6), New Guinea (1), Other Near Oceania (1)	AY289075M/Q1
DQ309854Kark	16075C 16129A 16144C 16148T 16221T 16223T 16241G 16261T 16265CtvA 16311C 16319A 16343G	1	Q1	New Guinea	
DQ309855Kark	16129A 16144C 16148T 16223T 16241G 16261T 16265CtvA 16311C 16343G	1	Q1	New Guinea	
U25374PNG	16093C 16129A 16144C 16148T 16209C 16221AtvC 16223T 16241G 16265CtvA 16311C 16343G	2	Q1	New Guinea	AY289081M/Q1
U25375PNG	16129A 16144C 16148T 16187T 16222T 16241G 16265CtvA 16311C 16343G	2	Q1	New Guinea	AY289082M/Q1
U25376PNG	16129A 16144C 16148T 16172C 16223T 16241G 16265CtvA 16311C 16343G	2	Q1	New Guinea	AF347003M/Q1
U25377PNG	16129A 16144C 16148T 16241G 16265CtvA 16311C 16343G 16362C	2	Q1	New Guinea	AY289085M/Q1
U25378PNG	16093C 16129A 16144C 16148T 16223T 16265CtvA 16311C 16343G 16362C	2	Q1	New Guinea	
U25379PNG	16129A 16144C 16148T 16223T 16241G 16265CtvA 16291T 16311C 16343G	2	Q1	New Guinea	
U47164Tahi2	16129A 16144C 16148T 16223T 16241G 16265CtvA 16293G 16311C 16343G	19	Q1	Remote Oceania	DQ372884M/Q1
U47165Tong1	16129A 16144C 16148T 16223T 16241G 16265CtvA 16311C 16343G	41	Q1	Indonesia (4), New Guinea (34), Remote Oceania (3)	AY289090, DQ372885 bothM/Q1
U47166Cook3	16129A 16144C 16148T 16223T 16241G 16265CtvA 16274A 16311C 16343G	3	Q1	Remote Oceania	
U47167Tong2	16129A 16144C 16148T 16223T 16241G 16265CtvA 16311C	2	Q1	Remote Oceania	
U47168PNG3	16129A 16148T 16223T 16241G 16265CtvA 16311C 16343G	3	Q1	New Guinea	
U47169Vanu1	16129A 16144C 16148T 16223T 16265CtvA 16311C 16343G	1	Q1	Remote Oceania	
U47170Vanu2	16129A 16144C 16148T 16265CtvA 16311C 16343G	4	Q1	Remote Oceania	DQ372880, DQ372882 bothM/Q1
AB119307Balo	16066G 16129A 16223T 16241G 16311C	1	Q2	Other Near Oceania	
AB119321Balo	16093C 16129A 16174T 16223T 16241G 16311C	1	Q2	Other Near Oceania	
AB119339Balo	16066G 16129A 16209C 16223T 16241G	6	Q2	Other Near Oceania	
AB119343Balo	16066G 16129A 16223T 16241G 16355T	2	Q2	New Guinea (1), Other Near Oceania (1)	
AB119367Tong	16066G 16129A 16223T 16241G 16294T 16352C	1	Q2	Remote Oceania	
AB119437Gidr	16093C 16129A 16209C 16223T 16241G 16311C	1	Q2	New Guinea	
AY956413Bism	16066G 16129A 16223T 16241G	2	Q2	Other Near Oceania	AY956412, AY956413M/Q2
AY956414Bism	16066G 16129A 16176T 16223T 16241G	1	Q2	Other Near Oceania	AY956414M/Q2
U47224Vanu1	16066G 16223T 16241G	1	Q2	Remote Oceania	
AB119444Gidr	16129A 16209C 16223T 16241G 16274A 16311C	5	Q3	New Guinea	
U25380PNG	16129A 16209C 16223T 16241G 16311C 16320T	2	Q3	New Guinea	AY289089M/Q3
U25381PNG	16129A 16209C 16223T 16241G 16311C	13	Q3	New Guinea	AY289078M/Q3
U25384PNG	16129A 16223T 16241G 16242AtvC 16311C	2	Q3	New Guinea	AY289079M/Q3
AB119285Balo	16278T	1		Other Near Oceania	
AB119331Balo	16066G 16129A 16223T 16239T 16325C	3		Other Near Oceania	
AB119340Balo	16126C 16129A 16223T 16297C	10		Borneo (1), Taiwan (1), Indonesia (3), Malaysia (2), Philippines (1), New Guinea (1), Other Near Oceania (1)	
AB119388Gidr	16066G 16067GtvC 16223T	1		New Guinea	
AB119400Gidr	16176T 16327T	1		New Guinea	
AB119418Gidr	16066G 16166G 16223T	2		New Guinea	
AB119420Gidr	16108T 16184T 16223T 16256T	4		New Guinea	
AB119422Gidr	16176T 16327T 16357C	1		New Guinea	

E6.1 cont. Haplotypes from HVR-I nt16065-nt16373 data set found in Oceania (page 5/6)

Haplotype	Differences to rCRS	n	Haplogroup	Regions	Entire mt genomes
AB119428Gidr	16231C 16271C 16286T 16309G	1		New Guinea	
AB119433Gidr	16176T 16189C 16191T 16357C	7		New Guinea	
AB119442Gidr	16093C 16184T 16223T 16256T	1		New Guinea	
AJ635002Iria	16145A 16243C 16278T 16293G 16319A 16343G	1		New Guinea	
AJ635003Iria	16184AtvC 16223T 16260T 16274A 16278T	3		New Guinea	
AJ635009Iria	16129A 16148T 16223T	1		New Guinea	
AJ635014Iria	16066G 16223T 16234T 16270T	1		New Guinea	
AJ635058Iria	16319A	1		New Guinea	
AJ635065Iria	16066G 16223T 16270T	5		New Guinea	
AJ635100Iria	16093C 16184AtvC 16223T 16278T	2		Borneo (1), New Guinea (1)	
AJ635104Iria	16066G 16172C 16173T 16290T 16298C 16232C	1		New Guinea	
AJ635108Iria	16066G 16086C 16223T 16278T	1		New Guinea	
AJ635111Iria	16223T 16260T 16274A 16278T	1		New Guinea	
AJ635117Iria	16176T 16320T 16357C	1		New Guinea	
AJ635122Iria	16066G 16223T 16278T	1		New Guinea	
AJ635128Iria	16291T 16311C 16327T 16335G	1		New Guinea	
AJ635129Iria	16172C 16176T 16320T 16357C	1		New Guinea	
AJ635135Iria	16066G 16223T	4		East Asia (1), New Guinea (3)	
AJ635149Iria	16176T 16256T 16320T 16357C	2		New Guinea	
AJ635150Iria	16093C 16172C 16173T 16223T 16247TtvA 16256T 16288C 16355T	2		New Guinea	
AY604130NZMaor	16126C 16292T 16294T	1		Remote Oceania	
AY604132NZMaor	16162G 16209C	1		Remote Oceania	
AY604134NZMaor	16093C 16224C 16311C	1		Remote Oceania	
AY604136NZMaor	16231C 16304C 16311C	1		Remote Oceania	
DQ309860Kark	16223T 16291T 16362C	93		East Asia (6), Borneo (4), Japan (1), Taiwan (23), Indonesia (46), Malaysia (5), Philippines (7), New Guinea (1)	AP008561M/D
DQ309862Kark	16126C 16129A 16223T	2		Indonesia (1), New Guinea (1)	
DQ309863Kark	16129A 16172C 16294T 16304C 16362C	34		East Asia (1), Borneo (3), Taiwan (6), Indonesia (22), Philippines (1), New Guinea (1)	
DQ309864Kark	16162G 16176T 16191T 16357C	1		New Guinea	
DQ309865Kark	16223T 16362C	104		East Asia (39), Borneo (5), Japan (30), Taiwan (11), Indonesia (13), Philippines (5), New Guinea (1)	AY289070M/M9/E, 13 JapaneseM/D
DQ309866Kark	16092C 16129A 16288C 16304G	1		New Guinea	
DQ309867Kark	16319A 16342C	1		New Guinea	
EF077389NZSolo	16086C 16209C 16223T 16299G	1		Remote Oceania	
U25390PNG	16184AtvC 16223T 16278T 16291T	1		New Guinea	
U47178Marq2	16172C 16304C	7		East Asia (2), Southeast Asia (2), Indonesia (1), Remote Oceania (2)	
U47179Tahi1	16172C 16189C 16304C	1		Remote Oceania	
U47180Tong1	16172C 16304C 16311C	2		Philippines (1), Remote Oceania (1)	
U47184Cook1	16263C	1		Remote Oceania	
U47185Cook1	16312G	1		Remote Oceania	
U47186Cook1	16356C	2		Remote Oceania	
U47187Cook1	16221T 16264T	1		Remote Oceania	
U47188Cook1	16210G 16256T 16269G	1		Remote Oceania	
U47190NZMaor1	16218T 16270T 16291T	1		Remote Oceania	

E6.1 cont. Haplotypes from HVR-I nt16065-nt16373 data set found in Oceania (page 6/6)

Haplotype	Differences to rCRS	n	Haplogroup	Regions	Entire mt genomes
U47191Tahi1	16223T 16290T 16319A 16362C	7		East Asia (4), Japan (2), Remote Oceania (1)	
U47192Tong1	16235G 16291T 16293G	1		Remote Oceania	
U47194Marq1	16172C 16223T	1		Remote Oceania	
U47195Marq1	16298C 16311C	1		Remote Oceania	
U47196Marq1	16126C 16153A 16294T	1		Remote Oceania	
U47197Marq1	16223T 16224C 16270T 16274A 16311C	1		Remote Oceania	
U47199Vanu1	16176T	1		Remote Oceania	
U47200Vanu1	no differences	16		East Asia (7), Indonesia (1), New Guinea (4), Remote Oceania(4)	DQ372870, DQ372872 bothN/R/P2
U47202Vanu1	16320T	1		Remote Oceania	
U47206Vanu2	16188T 16250T	2		Remote Oceania	
U47213Vanu1	16209C 16266T	1		Remote Oceania	
U47218Vanu1	16256T 16261T	1		Remote Oceania	
U47232Vanu1	16169T 16176T 16209C	1		Remote Oceania	
U47233Vanu1	16172C 16223T 16264T	1		Remote Oceania	
U47248Vanu1	16290T 16298C 16262C	1		Remote Oceania	
U47256Mars1	16214T 16223T 16295T 16362C	1		Remote Oceania	
U47261Vanu1	16066G 16129A 16168T 16172C 16223T	1		Remote Oceania	
U47264Vanu1	16172C 16223T 16264T 16311C 16320T	1		Remote Oceania	

E6.2 Haplotype details for HVR-I nt16189-nt16373 data set (page 1/4)

This table contains details of the 199 haplotypes from the HVR-I nt16189-nt16370 data set. It is ordered firstly by assigned haplogroup, and then by haplotype name. The differences to the rCRS for each haplotype were obtained using Sequencher™ (Gene Codes Corporation).

Haplotype	Differences to rCRS	n	Haplogroup	Regions
DQ137409Bism	T16189C C16223T T16311C	3	M/M27a	Other Near Oceania
DQ137411Bour	T16189C C16223T T16311C C16320T	2	M/M27a	Other Near Oceania
DQ137406Bism	C16223T C16301T T16304C	2	M/M27c	Other Near Oceania
DQ137398Bism	C16223T A16318tvC T16362C	1	M/M28b	Other Near Oceania
DQ137399Bism	C16223T C16291T A16318tvT T16362C	1	M/M28b	Other Near Oceania
AB119383Tong	C16223T C16295T	2	M/M7c	Marshall Islands (1), Tonga (1)
AF285729pohn	C16223T C16295T T16311C T16362C	1	M/M7c	Pohnpei
U47176Mars2	C16223T C16295T T16362C	15	M/M7c	Nauru (2), Kosrae (4), Kiribati (1), Pohnpei (5), Marshall Islands (3)
U47256Mars1	C16214T C16223T C16295T T16362C	5	M/M7c	Kiribati (4), Marshall Islands (1)
AB119321Balo	C16223T A16241G T16311C	5	M/Q	New Guinea (3), Other Near Oceania (2)
AB119339Balo	T16209C C16223T A16241G	6	M/Q	Other Near Oceania
AB119367Tong	C16223T A16241G C16294T T16352C	1	M/Q	Tonga
AB119444Gidr	T16209C C16223T A16241G G16274A T16311C	5	M/Q	New Guinea
AF066319pohn	C16223T A16241G C16355T	3	M/Q	New Guinea (1), Other Near Oceania (1), Pohnpei (1)
AF066471vanu	C16223T A16241G	11	M/Q	New Guinea (1), Other Near Oceania (3), Vanuatu (7)
AJ635116Iria	C16214T C16223T T16231C A16241G T16311C	1	M/Q	New Guinea
AJ635137Iria	C16214T C16223T A16241G T16311C	2	M/Q	New Guinea
AJ635159Iria	C16223T A16241G A16265tvC C16270T T16311C	2	M/Q	New Guinea
U25380PNG	T16209C C16223T A16241G T16311C C16320T	2	M/Q	New Guinea
U25381PNG	T16209C C16223T A16241G T16311C	14	M/Q	New Guinea
U25384PNG	C16223T A16241G C16242tvA T16311C	2	M/Q	New Guinea
U47167Tong2	C16223T A16241G A16265tvC T16311C	2	M/Q	Tonga
U47193tong1	T16189C C16223T A16241G	1	M/Q	Tonga
U47243Vanu1	C16223T A16241G G16319A	1	M/Q	Vanuatu
U47271Vanu1	C16223T C16234T A16241G A16269G C16292T T16297C T16311C T16362C	1	M/Q	Vanuatu
AB119306Balo	C16223T C16261T A16265tvC T16311C A16343G	2	M/Q1	New Guinea (1), Other Near Oceania (1)
AB119336Balo	C16223T A16241G C16259tvA A16265tvC A16300G T16311C A16343G	5	M/Q1	Other Near Oceania
AF066448fiji	C16223T A16241G A16265tvC T16311C A16343G T16362C	1	M/Q1	Fiji
AF066477vanu	C16222T A16241G A16265tvC T16311C A16343G	14	M/Q1	New Guinea (13) and Vanuatu (1)
AF285735samo	C16223T A16241G A16265tvC T16311C A16343G	72	M/Q1	New Guinea (68), Tonga (1), Samoa (3)
AJ635000Iria	C16222T A16241G T16243C A16265tvC T16311C A16343G	1	M/Q1	New Guinea
AJ635006Iria	C16223T A16241G A16265tvC G16273A G16274A T16304C T16311C A16343G	1	M/Q1	New Guinea
AJ635007Iria	C16218T C16223T A16241G T16243C A16265tvC T16311C A16343G	1	M/Q1	New Guinea
AJ635008Iria	C16223T C16234T A16241G A16265tvC T16311C A16343G	1	M/Q1	New Guinea
AJ635011Iria	C16223T T16263C A16265tvC T16311C A16343G	1	M/Q1	New Guinea
AJ635013Iria	C16223T C16261T A16265tvC C16291T T16298C A16299G T16311C A16343G	2	M/Q1	New Guinea
AJ635043Iria	C16223T A16241G A16265tvC C16292T T16311C A16343G	2	M/Q1	New Guinea
AJ635045Iria	T16209C C16223T A16241G A16265tvC T16311C A16343G	4	M/Q1	New Guinea
AJ635051Iria	C16223T A16241G A16265tvC C16290T T16311C A16343G	2	M/Q1	New Guinea
AJ635057Iria	C16223T C16261T A16265tvC C16291T T16298C T16311C A16343G	8	M/Q1	New Guinea
AJ635060Iria	C16223T A16227G A16241G A16265tvC T16311C A16343G	2	M/Q1	New Guinea
AJ635073Iria	G16213A C16223T A16241G A16265tvC T16311C A16343G	3	M/Q1	New Guinea
AJ635099Iria	C16223T A16241G A16265tvC C16270T T16311C A16343G	1	M/Q1	New Guinea
AJ635112Iria	C16223T T16224C A16241G A16265tvC T16311C A16343G	3	M/Q1	New Guinea
AJ635114Iria	C16222T A16241G A16265tvC T16304C T16311C A16343G	1	M/Q1	New Guinea
AJ635118Iria	C16223T A16265tvC A16305G T16311C A16343G	1	M/Q1	New Guinea
AJ635153Iria	C16223T A16241G T16243C A16265tvC T16311C A16343G	3	M/Q1	New Guinea

E6.2 cont. Haplotype details for HVR-I nt16189-nt16373 data set (page 2/4)

Haplotype	Differences to rCRS	n	Haplogroup	Regions
AY289075Bour	A16241G A16265tvC T16311C A16343G	3	M/Q1	New Guinea (2), Other Near Oceania (1)
DQ309854Kark	C16221T C16223T A16241G C16261T A16265tvC T16311C G16319A A16343G	1	M/Q1	New Guinea
DQ309855Kark	C16223T A16241G C16261T A16265tvC T16311C A16343G	1	M/Q1	New Guinea
U25374PNG	T16209C C16221tvA C16223T A16241G A16265tvC T16311C A16343G	2	M/Q1	New Guinea
U25376PNG	C16223T A16241G G16255A A16265tvC T16311C A16343G	2	M/Q1	New Guinea
U25377PNG	A16241G A16265tvC T16311C A16343G T16362C	2	M/Q1	New Guinea
U25378PNG	C16223T A16265tvC T16311C A16343G T16362C	2	M/Q1	New Guinea
U25379PNG	C16223T A16241G A16265tvC C16291T T16311C A16343G	2	M/Q1	New Guinea
U47164NZMaor1	C16223T A16241G A16265tvC A16293G T16311C A16343G	16	M/Q1	Samoa (1), Other East Polynesia (2), Cook Islands (12), New Zealand (1)
U47166Cook3	C16223T A16241G A16265tvC G16274A T16311C A16343G	5	M/Q1	New Guinea (2), Cook Islands (3)
U47169Vanu1	C16223T A16265tvC T16311C A16343G	12	M/Q1	New Guinea (1), Yap (9), Pohnpei (1), Fiji (1)
U47170Vanu2	A16265tvC T16311C A16343G	4	M/Q1	Vanuatu
AB119346Haap	T16189C T16217C C16261 T16311C	2	N/R/B4a	Fiji (1), Tonga (1)
AF066241kosr	T16189C T16217C C16278T T16311C	1	N/R/B4a	Kosrae
AF066260mars	T16189C T16217C C16261T C16301T	1	N/R/B4a	Marshall Islands
AF066270mars	T16189C T16217C C16261T T16362C	1	N/R/B4a	Marshall Islands
AF066322pohn	T16189C T16217C C16278T	15	N/R/B4a	Yap (15), Nauru (2), Kiribati (1), Pohnpei (1)
AF066561pala	T16189C T16217C C16261T G16319A	3	N/R/B4a	Palau
AF285719pala	T16189C T16217C A16289G	11	N/R/B4a	Palau
U47148kapa8	T16189C T16217C	25	N/R/B4a	Other Near Oceania (1), Yap (1), Palau (3), Kiribati (1), Pohnpei (2), Kapingamarangi (17)
U47150NZMaor2	T16189C T16217C C16261T	114	N/R/B4a	New Guinea (5), Yap (31), Palau (9), Marianas (1), Nauru (2), Kosrae (2), Kiribati (7), Pohnpei (1), Marshall Islands (9), Kapingamarangi (13), Vanuatu (5), Tonga (6), Samoa (5), Marquesas (1), Other east Polynesia (2), Cook Islands (10), New Zealand (5)
U47154cook1	T16189C T16217C C16261T A16312G	1	N/R/B4a	Cook Islands
AB119299Balo	T16189C T16217C A16246G A16247G	1	N/R/B4a1a1PM	Other Near Oceania
AB119302Balo	T16189C T16217C A16247G C16261T A16293G	3	N/R/B4a1a1PM	Other Near Oceania
AB119337Balo	T16189C T16217C C16242tvA A16247G C16261T	2	N/R/B4a1a1PM	Other Near Oceania
AB119345Haap	T16189C T16217C C16218T A16247G C16261T	1	N/R/B4a1a1PM	Tonga
AB119352Haap	T16189C T16217C A16247G C16248T C16261T	5	N/R/B4a1a1PM	Nauru (4), Tonga (1)
AB119377Tong	T16189C T16217C A16247G C16261T C16278T	1	N/R/B4a1a1PM	Tonga
AF066106yap	T16189C T16217C A16247G C16261T G16274A	1	N/R/B4a1a1PM	Yap
AF066209yap	T16189C T16217C A16247G C16261T G16319A	24	N/R/B4a1a1PM	Yap
AF066235kosr	T16189C T16217C C16218T A16247G C16261T C16287T	1	N/R/B4a1a1PM	Kosrae
AF066279naur	T16189C T16217C A16247G C16248T C16259T C16261T	1	N/R/B4a1a1PM	Nauru
AF066293naur	T16189C T16217C A16247G C16261T C16295T	1	N/R/B4a1a1PM	Nauru
AF066584pala	T16189C T16217C A16247G C16262T	2	N/R/B4a1a1PM	Palau
AF285734samo	T16189C T16217C A16247G C16261T C16262T	1	N/R/B4a1a1PM	Samoa
AF285737samo	T16189C T16217C A16247G C16294T	1	N/R/B4a1a1PM	Samoa
AF285739samo	T16189C T16217C A16247G C16261T T16263C	3	N/R/B4a1a1PM	Samoa
AF347007Samo	T16189C T16217C A16247G C16261T T16324C	1	N/R/B4a1a1PM	Samoa
AY289069Cook	T16189C T16217C A16247G C16261T A16317G	1	N/R/B4a1a1PM	Cook Islands
AY289094Samo	T16189C T16217C A16247G C16261T T16352C	1	N/R/B4a1a1PM	Samoa
AY289102Tong	T16189C T16217C A16247G C16261T C16360T	1	N/R/B4a1a1PM	Tonga
AY604140NZMaor	T16189C T16217C C16242T A16247G C16261T	1	N/R/B4a1a1PM	New Zealand
DQ309851Kark	T16189C A16247G C16261T T16357C	2	N/R/B4a1a1PM	New Guinea
DQ309852Kark	T16189C A16247G C16261T	2	N/R/B4a1a1PM	New Guinea
EF077370NZTong	T16189C T16217C A16247G C16261T T16311C	8	N/R/B4a1a1PM	Other Near Oceania (3), Fiji (1), Tonga (4)
EF077372NZSamo	T16189C T16217C A16247G T16249C C16261T	1	N/R/B4a1a1PM	Samoa
U47155nzmaor20	T16189C T16217C A16247G C16261T	340	N/R/B4a1a1PM	Papua New Guinea (2), Karkar is (1), Balopa Is (22), Yap (53), Palau (33), Marianas (4), Nauru (15), Kosrae (4), Kiribati (5), Pohnpei (6), Marshall Is (3), Kapingamarangi (3), Vanuatu (8), Fiji (6), Tonga (35), Samoa (30), Marquesas (12), Other east Polynesia (12), Niue (2), Cook Is (54), New Zealand (32)
U47156mars1	T16189C T16217C A16247G C16261T C16287T	21	N/R/B4a1a1PM	Kosrae (9), Marshall Islands (12)
U47157cook1	T16189C T16217C A16247G C16261T C16291T	3	N/R/B4a1a1PM	Tonga (1), Cook Islands (2)

E6.2 cont. Haplotype details for HVR-I nt16189-nt16373 data set (page 3/4)

Haplotype	Differences to rCRS	n	Haplogroup	Regions
U47158samo1	T16189C T16217C A16247G C16261T C16354tvA	1	N/R/B4a1a1PM	Samoa
U47159samo1	T16189C T16217C A16247G C16261T G16274tvC	1	N/R/B4a1a1PM	Samoa
U47160samo1	T16189C T16217C A16247G C16261T T16342C	3	N/R/B4a1a1PM	Tonga (1), Samoa (2)
U47162samo1	T16189C T16217C A16247G C16261T T16362C	3	N/R/B4a1a1PM	Pohnpei (1), Tonga (1), Samoa (1)
U47163cook1	T16189C T16217C A16247G	8	N/R/B4a1a1PM	Other Near Oceani (6), Other east Polynesia (1), Cook Islands (1)
AB119440Gidr	T16189C C16266T C16270T T16357C	1	N/R/P1	New Guinea
AF066120yap	G16255A C16259tvG C16266T C16270T T16357C	1	N/R/P1	Yap
AF066226kiri	C16259tvG C16266T C16270T G16319A T16357C	1	N/R/P1	Kiribati
AF066301naur	C16259tvG C16266T C16270T T16357C	3	N/R/P1	Yap (2), Nauru (1)
AF066459vanu	C16266T C16291T T16311C T16357C	2	N/R/P1	Vanuatu
AF066462vanu	T16263C C16266T C16291T T16311C T16357C	1	N/R/P1	Vanuatu
AF066636pala	C16266T C16270T T16311C T16357C	2	N/R/P1	Palau
AF285714pala	C16266T C16270T C16301T T16357C	2	N/R/P1	Palau
AF347004PNG	C16266T C16294T T16357C	1	N/R/P1	New Guinea
AF347005PNG	T16189C C16223T A16235G C16257T C16266T C16270T C16354T T16357C	1	N/R/P1	New Guinea
AJ635041Iria	G16213A C16266T C16270T T16357C	1	N/R/P1	New Guinea
AJ635067Iria	C16225T C16266T C16270T A16318G T16357C	2	N/R/P1	New Guinea
AJ635142Iria	C16223T C16262T C16266T C16270T T16311C T16357C	6	N/R/P1	New Guinea
AJ635154Iria	C16221T C16266T C16270T T16311C T16357C	6	N/R/P1	New Guinea
AY289092PNG	C16266T C16291T C16292T A16335G T16357C	1	N/R/P1	New Guinea
U25385PNG	C16257T C16266T C16270T C16354T T16357C	2	N/R/P1	New Guinea
U25386PNG	A16235G C16257T C16264T C16266T C16270T T16304C C16354T T16357C	1	N/R/P1	New Guinea
U25387PNG	A16235G C16266T T16357C	2	N/R/P1	New Guinea
U25388PNG	C16266T C16287T C16294T T16357C	2	N/R/P1	New Guinea
U47220Vanu1	C16266T T16357C	11	N/R/P1	New Guinea (8), Other Near Oceania (2), Vanuatu (1)
U47254Vanu3	C16266T C16270T T16357C	51	N/R/P1	New Guinea (14), Palau (34), Vanuatu (3)
U47265Vanu1	T16209C C16266T G16274A T16357C	1	N/R/P1	Vanuatu
U47269Vanu1	C16250T C16266T C16291T T16311C T16357C	1	N/R/P1	Vanuatu
AB119285Balo	C16278T	1		Other Near Oceania
AB119331Balo	C16223T C16239T T16325C	3		Other Near Oceania
AB119340Balo	C16223T T16297C	2		New Guinea (1), Other Near Oceania (1)
AB119365Tong	C16223T C16366T T16362C	3		Vanuatu (1), Tonga (2)
AB119400Gidr	T16325C	1		New Guinea
AB119422Gidr	C16327T T16357C	1		New Guinea
AB119428Gidr	T16231C T16271C C16286T A16309G	1		New Guinea
AB119442Gidr	C16223T C16256T	5		New Guinea
AF066390yap	A16219G C16223T	4		Yap
AF066474vanu	T16357C	9		New Guinea (8), Vanuatu (1)
AF066478vanu	C16223T T16362C	27		New Guinea (1), Other Near Oceania (1), Yap (4), Marianas (9), Vanuatu (11)
AF066517mari	T16209C C16223T A16302G T16362C	3		Marianas
AF066521mari	C16256T T16304C A16335G	1		Marianas
AF066537pala	T16189C C16287T	1		Palau
AF066542pala	T16189C	1		Palau
AF066553pala	C16223T T16297C C16354T	1		Palau
AF066587pala	C16223T T16271C T16362C	1		Palau
AF066588pala	C16223T G16274A T16311C A16317G T16362C	5		Palau
AF066595pala	C16270T T16357C	1		Palau
AF066639pala	T16189C T16243C	10		Yap (5), Palau (5)
AF066640pala	C16294T T16304C T16362C	3		New Guinea (1), Palau (2)
AF285672mars	T16189C C16270T	1		Marshall Islands
AF285753yap	T16362C	5		Yap
AJ635002Iria	T16243C C16278T A16293G G16319A A16343G	1		New Guinea
AJ635014Iria	C16223T C16234T C16270T	1		New Guinea
AJ635058Iria	G16319A	1		New Guinea
AJ635065Iria	C16223T C16270T	5		New Guinea
AJ635111Iria	C16223T C16260T G16274A C16278T	4		New Guinea
AJ635122Iria	C16223T C16278T	3		New Guinea
AJ635128Iria	C16291T T16311C T16325C A16335G	1		New Guinea
AJ635129Iria	C16320T T16357C	2		New Guinea
AJ635149Iria	C16256T C16320T T16357C	2		New Guinea
AJ635150Iria	C16223T C16256T T16288C C16355T	2		New Guinea
AY289088PNG	C16278T T16357C	2		New Guinea

E6.2 cont. Haplotype details for HVR-I nt16189-nt16373 data set (page 4/4)

Haplotype	Differences to rCRS	n	Haplogroup	Regions
AY604130NZMaor	C16292T C16294T	1		New Zealand
AY604132NZMaor	T16209C	2		Vanuatu (1), New Zealand (1)
AY604133NZMaor		13		New Guinea (8), Vanuatu (4), New Zealand (1)
AY604134NZMaor	T16224C T16311C	1		New Zealand
AY604135NZMaor	T16356C	2		Cook Islands (1), New Zealand (1)
AY604136NZMaor	T16231C T16304C T16311C	1		New Zealand
DQ137400Bism	C16223T C16320T T16362C	1		Other Near Oceania
DQ137404Bism	T16209C A16299G	3		Other Near Oceania
DQ309860Kark	C16223T C16291T T16362C	1		New Guinea
DQ309866Kark	T16288C T16304tvG	1		New Guinea
DQ309867Kark	G16319A T16342C	1		New Guinea
EF077389NZSolo	T16209C C16223T A16299G	1		Other Near Oceania
U25390PNG	C16223T C16278T C16291T	1		New Guinea
U47171Vanu2	C16223T T16311C T16362C	5		Nauru (2), Vanuatu (3)
U47173Vanu1	C16223T C16355T T16362C	1		Vanuatu
U47174Vanu1	C16223T C16355T C16365T T16362C	1		Vanuatu
U47178Marq2	T16304C	2		Marquesas
U47179Tahi1	T16189C T16304C	1		Other East Polynesia
U47180Tong1	T16304C T16311C	1		Tonga
U47184Cook1	T16263C	1		Cook Islands
U47185Cook1	A16312G	1		Cook Islands
U47187Cook1	C16221T C16264T	1		Cook Islands
U47188Cook1	A16210G C16256T A16269G	1		Cook Islands
U47189Cook1	C16223T T16298C T16325C	1		Cook Islands
U47190NZMaor1	C16218T C16270T C16291T	1		New Zealand
U47191Tahi1	C16223T C16290T G16319A T16362C	1		Other East Polynesia
U47192Tong1	A16235G C16291T A16293G	1		Tonga
U47194Marq1	C16223T	10		New Guinea (8), Vanuatu (1), Marquesas (1)
U47195Marq1	T16298C T16311C	1		Marquesas
U47196Marq1	C16294T	1		Marquesas
U47197Marq1	C16223T T16224C C16270T G16274A T16311C	1		Marquesas
U47202Vanu1	C16320T	1		Vanuatu
U47206Vanu2	C16250T	2		Vanuatu
U47213Vanu1	T16209C C16266T	1		Vanuatu
U47218Vanu1	C16256T C16261T	1		Vanuatu
U47233Vanu1	C16223T C16264T	1		Vanuatu
U47236vanu1	T16189C C16223T C16294T	1		Vanuatu
U47240vanu1	T16189C C16290tvA	1		Vanuatu
U47248Vanu1	C16290T T16298C T16362C	2		New Guinea (1), Vanuatu (1)
U47253Vanu2	T16209C T16357C	9		New Guinea (6), Fiji (1), Vanuatu (2)
U47264Vanu1	C16223T C16264T T16311C C16320T	1		Vanuatu